GENOME-WIDE PATTERNS OF POPULATION STRUCTURE AND ANCESTRY AMONG CONTINENTAL AND ADMIXED POPULATIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Katarzyna Bryc

January 2011

GENOME-WIDE PATTERNS OF POPULATION STRUCTURE AND ANCESTRY AMONG CONTINENTAL AND ADMIXED POPULATIONS

Katarzyna Bryc, Ph.D.

Cornell University 2011

Population genetics seeks to use genetic data to illuminate patterns of human diversity, investigate how populations are related, and to provide insights into population history, such as migrations events and population sizes. Furthermore, an understanding of population genetics is necessary to disentangle population structure from genetic associations with traits, to learn how genes affect phenotype or to perform disease association mapping.

I use high-density single nucleotide polyphorphism (SNP) data to examine population structure in humans among several world-wide populations. I show that principal components analysis (PCA) and *STRUCTURE*, a bayesian clustering method, are able to resolve structure both among continents as well as illuminate substructure within Europe, South Asia, and East Asia. In an analysis of 12 West African populations, I demonstrate that population structure within the West African samples reflects linguistic relationships and geographical distances, and also shows signals of the Bantu expansion.

I proceed to focus on several questions involving populations of mixed ancestry, or admixed populations. First, I introduce a new method for inferring individual ancestry along the genome, or "local ancestry". This method leverages principal component analysis to allow computationally efficient ancestry estimation using high-density SNP data. I apply this method to a sample of African Americans and witness a large range of ancestry proportions across in-

dividuals in this panel. I find that the African Americans have a greater propotion of African ancestry on the X chromosome versus the autosomes, consistent with a greater female African and male European ancestry contribution. Since previous studies have suggested a West African ancestral population of African Americans, I use estimates of African and European segments of the genome to examine which of 12 West African populations is closest to the African ancestral population. I find that, consistent with the West African results of previous studies and historical records, the African regions of African American genomes show the lowest genetic divergence to West African populations Igbo, Brong, and Yoruba, which are non-Bantu Niger-Kordofanian speaking populations.

Hispanic/Latino (HL) populations possess a complex genetic structure reflecting recent admixture among Native American, European, and West African populations. I estimate ancestry among five Hispanic/Latino populations (Mexico, Ecuador, Colombia, Puerto Rico, and Dominican Republic) and illuminate patterns of ancestry among populations. These differences among HL populations reflect geographic proximity to slave trade routes and ports, European colonizations, and historical migrations. I show a consistent sex bias in ancestry proportions across all five HL populations with higher Native American and lower European ancestry on the X chromosome compared to the autosomes. The ancestry difference on the X versus the autosomes suggests a greater Native American female and European male ancestry contribution bias in all five HL populations, and is further supported by Y chromosome and mitochondrial DNA haplotyping. Lastly, I discuss challenges in identifying the closest Native American ancestral population to the HL populations, such as poor Native American population sampling or substructure within the Americas. However, I am able to show that the Nahua (for Meso-American populations) and the

Quechua (for South American populations) are the two populations least differentiated from the Native American segments of the HL individuals.

BIOGRAPHICAL SKETCH

Katarzyna (Kasia) Bryc was born on August 22, 1983 in Lublin, Poland to parents Wlodzimierz (Wlodek) and Grazyna Bryc. After emigrating in 1985, she grew up in and around Cincinnati, Ohio. She first became interested in becoming an archaeologist after her parents took her to a natural history museum at the age of five. This unusual career aspiration led to stumping her first grade teacher when she asked for help spelling her desired future profession. Later in grade school, she first discovered that she had a knack for problem solving in the seventh grade when she surpassed everyone in her entire middle school on a school-wide math competition.

In high school, she continued to focus her energy studying math and science, though her preference for visual geometry over formulaic algebra showed by her taking first place in the state of Ohio for Geometry in a national math competition in 1999. After graduating from St. Ursula Academy in 2001 earning the school's highest honors in physics and computer science, she went on to study Mathematical and Computational Science at Stanford University. At Stanford, she grew diverse interests both in and outside the classroom. She enjoyed courses in computer science, probability and statistics, optimization, and theoretical mathematics, as well as Russian language, human osteology, and geology. Much to her father's dismay, she went on to join the NCAA division I varsity fencing team, became president of the Stanford Outdoor club, taught Stanford's windsurfing class, and lead incoming students on Stanford Pre-Orientation Trips. Though she did a summer research internship in 2004 with Prof. Susan Holmes where she worked with a biotechnology company and studied microarray data, it was in her anthropology class with Prof. Joanna

Mountain that she discovered her passion for population genetics, which combined her interest in thinking about evolution and her skills in statistics.

After earning her B.S. from Stanford in 2005, she worked in the Mountain lab studying African population genetics, an interest she would later return to in graduate school. Following the internship, she backpacked for a few months through Europe, but returned to her hometown to work with Dr. Todd Nick at Cincinnati Children's Hospital Medical Center as a biostatistics research assistant, where she was introduced to the clinical side of epidemiology and pharmacogenetics. Kasia started her graduate work in the field of Biometry at Cornell University in the Fall of 2006, and immediately joined the exciting lab of Dr. Carlos D. Bustamante, and has worked on the population genetics of several species for the past four years under his direction.

She still hopes to someday become an archaeologist, though not until after her retirement.

For my parents, who have supported me beyond measure.

ACKNOWLEDGEMENTS

There are many people I would like to thank who have made this work possible.

My advisor Carlos Bustamante, who has enthusiastically supported my research, provided unending data and project directions, ideas and creativity, and who gave me an unparalleled introduction into the field of population genetics. I personally am grateful for your advice both on research questions and on how to navigate the world of academia, and I am honored to have had the opportunity to learn from you as a scientist and scholar.

I would also like to thank my committee members, Andy Clark and Alon Keinan. I am grateful for your encouragement in my research, and for your helpful directions and comments that always kept me on track. I would also like to thank my many colleagues who allowed me the honor to collaborate with them as I navigated the new waters of large projects: John Novembre, Matthew Nelson, Adam Auton, Sarah Tishkoff, and Harry Ostrer and many others. Lastly, it is important to acknowledge the many wonderful former Bustamante labmates I've had over the years, who provided day to day support and help: Jeremiah Degenhardt, Abra Brisbin, Amit Indap, Andy Reynolds, Adam Auton, and so many others. Thank you for answering silly scripting questions, lending an ear when needed, and inviting me for indispensible coffee breaks.

Lastly, my warmest thanks go to my parents who have supported me for so many years, and who have always been the pillar I could lean on when things become difficult. Your belief in me made it possible for me to overcome the many hurdles along the way.

TABLE OF CONTENTS

Bibliography **162**

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

THE POPULATION REFERENCE SAMPLE (POPRES): A RESOURCE FOR POPULATION, DISEASE, AND PHARMACOLOGICAL GENETICS RESEARCH*

*Originally published as: Nelson, M. R., K. Bryc, K. S. King, A. Indap, A. R. Boyko, J. Novembre, L. P. Briley, Y. Maruyama, D. M. Waterworth, G. Waeber, P. Vollenweider, J. R. Oksenberg, S. L. Hauser, H. A. Stirnadel, J. S. Kooner, J. C. Chambers, B. Jones, V. Mooser, C. D. Bustamante, A. D. Roses, D. K. Burns, M. G. Ehm, and E. H. Lai (2008). The Population Reference Sample: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet*, 83(3):347-358.

1.1 Abstract

Technological and scientific advances, stemming in large part from the Human Genome and HapMap projects, have made large-scale, genome-wide investigations feasible and cost-effective. These advances have the potential to dramatically impact drug discovery and development by identifying genetic factors that contribute to variation in disease risk as well as drug pharmacokinetics, treatment efficacy, and adverse drug reactions. In spite of the technological advancements, successful application in biomedical research would be limited without access to suitable sample collections. To facilitate exploratory genetic research, we have assembled a DNA resource from a large number of subjects participating in multiple studies throughout the world. This growing resource was initially genotyped using a commercially available genome-wide 500,000 SNP panel. This project includes nearly 6,000 subjects of African American, East Asian, South Asian, Mexican, and European origin. Seven informative axes of variation identified via principal component analysis (PCA) of these data confirm the overall integrity of the data and highlight important features of the genetic structure of diverse populations. The potential value of such extensively genotyped collections is illustrated by selecting genetically matched population controls in a genome-wide analysis of abacavir-associated hypersensitivity reaction. We find that matching based on country of origin, identity-by-state distance, and multidimensional PCA do similarly well to control the type I error rate. The genotype and demographic data from this reference sample are freely available through the NCBI database of Genotypes and Phenotypes (dbGaP).

1.2 Introduction

Our capacity to measure human genetic variation and apply it to address scientific questions related to evolution [Lohmueller et al., 2008], population structure [Jakobsson et al., 2008, Li et al., 2008], and interindividual phenotypic variation [WTCCC, 2007] is expanding at an increasing rate. At least as important as the technologies to measure genetic variation are the availability of suitable samples and their descriptive data. In the past, the resources to conduct large-scale genetic investigations have been restricted to a relatively small number of well-funded academic and commercial groups, limiting the access to the raw data. However, recent changes in attitudes in the scientific community, ethical review boards, and the policies of funding agencies are leading to greater openness in sharing genetic data with the intent to improve opportunities for discovery through their creative use and careful integration [Manolio et al., 2007, Mailman et al., 2007].

In 2005, GlaxoSmithKline initiated the Population Reference Sample (POPRES) project with the goal of bringing together a DNA sample set that would be extensively genotyped in order to support a variety of efforts related to pharmacogenetic research. We found that the application of pharmacogenetics research associated with drug development could be hampered by 1) lack of readily available population controls for adequately powered study designs, 2) high costs of conducting highly exploratory genome-wide studies, 3) extended study timelines that may not meet clinical development needs, and 4) lack of samples representative of the multinational patient populations from which the prevalence of pharmacogenetically relevant polymorphisms can be estimated. The POPRES project was carried out to begin addressing these issues, with the

further objective of making the resulting genotypic and demographic data publicly available to help drive development in the external genetics research community.

There are many projects, especially in pharmacogenetics, wherein the sample collection is focused on the acquisition of cases. One important example is the identification and collection of cases with adverse drug reactions (ADRs) through post-marketing surveillance. In these situations, the acquisition or selection of a suitable set of controls can add a substantial burden to the experimental process. Having a large collection of DNA samples previously scrutinized and genetically characterized would facilitate the search for genetic risk factors. This is particularly true if the case samples to be matched with the cohorts are not of Northern European origin, the background of most genome-wide studies published to date and publicly available. Availability of key demographic, phenotypic, and clinical data for the selected subjects would enhance their application.

Investigations into genetic risk factors underlying ADRs are highly exploratory, as there is generally little a priori evidence to support a genetic hypothesis. The availability of population cohorts with existing genotype data that could be matched to the cases substantially lowers the cost and time to conduct this research and could facilitate exploratory efforts. For ADRs that have relatively low frequency, there is little power lost in the use of population, versus drug-treated, clinically-matched chohort [Nelson et al., 2008]. A large resource of genotyped cohort would also allow for more careful matching of what can be genetically diverse cases to controls on the basis of their patterns of genetic variation [Luca et al., 2008].

Many pharmacogenetic studies utilize samples collected in clinical trials, which are becoming increasingly global and diverse in their origin [Thiers et al., 2008]. Therefore, in addition to the value of genome-wide genotype data for exploratory scans, the availability of DNA for the subjects included in the POPRES initiative allows for measurement of variants that are of particular interest to pharmacogenetic research, as well as estimation of their population-specific relative frequencies. This can be useful for predicting population-specific ADR risks or possible variability in drug response. Furthermore, population genetic studies of more diverse samples, such as POPRES, provide important information about the similarity or differentiation of these populations [Rosenberg et al., 2002], informing future study designs and interpretation.

The availability of a densely genotyped population reference sample will increase opportunities for many areas of genetics research, by us and others, by providing a well-characterized readily available set of samples representative of the populations of interest from which to draw controls and estimate population parameters of interest. Furthermore, such resources will foster development of statistical methods and analysis strategies and provide a resource for innovative population genetics research. In this paper, we describe the collections currently comprising 5,886 POPRES subjects, genotyping and analysis methods used in preparing the data being provided to the public domain, and selected data analysis results. Lastly, we present an example application matching controls to a small set of ADR cases.

1.3 Materials and Methods

The subjects included in the POPRES initiative are derived from ten collections. Each collection is briefly described. Where available, see accompanying references for further collection details. All subjects included in this study were either collected in an anonymous fashion, or have been multiply coded by the collecting institution as well as the POPRES data managers (see [on Harmonization, 2008] for definitions).

UCSF African Americans African American subjects were recruited across the United States to serve as controls for studies of multiple sclerosis (MS) genetic susceptibility conducted at the University of California San Francisco [Oksenberg et al., 2004]. In general, individuals were invited to participate in the study by the probands and constitute primarily spouses or friends of MS patients. In addition to the ability to give consent and willingness to participate, inclusion criteria included male and female gender, age > 16 years old, no personal or familial history of MS, and no history of autoimmunity. Exclusion criteria included chronic diseases and recreational drug use. All study participants were self-reported African-Americans, but European ancestry was documented based on genotyping results of 186 informative SNPs [Patterson et al., 2004].

Healthy Japanese Cohort Participants were recruited through the James Lance GlaxoSmithKline Medicines Research Unit in Sydney, Australia. Eligibility criteria included self-described Japanese ethnic background, older than 20 years of age, and free from chronic disease. Blood samples were collected in an anonymous fashion, i.e. no identifiers were associated with the biological

sample that could associate it back with the participant. Sex is the only personal information recorded for each subject.

Healthy Taiwanese Cohort Participants were recruited through the Tri-Service General Hospital in Taipei, Taiwan. Eligibility criteria included self-described ethnicity as Han Chinese, at least 20 years of age, and free from chronic disease. Blood samples were collected in an anonymous fashion. Sex is the only personal information recorded for each subject.

Healthy Mexican Cohort Participants were recruited through a hospital-based clinic in Guadalajara, Mexico. Eligibility criteria included self-described ethnicity as Mexican/Hispanic, at least 18 years of age, and free from chronic disease. Blood samples were collected in an anonymous fashion. Sex is the only personal information recorded for each subject.

Healthy Caucasian Cohort Participants were recruited through 1) the Royal Adelaide Hospital in Adelaide, Australia; 2) Duke University, North Carolina, USA; and 3) the University of Ottawa Heart Institute, Ottawa, Canada. Inclusion criteria included self-described ethnicity as Caucasian, at least 18 years old and healthy. Here, healthy is defined as individuals who are free from clinical cardiac, pulmonary, gastrointestinal, hepatic, renal, hematological, neurological and psychiatric disease as determined by history, physical examination or screening investigations. Blood samples were collected in an anonymous fashion. Sex is the only personal information recorded for each subject.

London Life Sciences Population (LOLIPOP) Study The LOLIPOP study is a population based study of Indian Asians and European whites aged 3575 years, identified from the lists of 58 general practitioners in West London [Kooner et al., 2008]. To date, 938 Northern Europeans and 431 Indian Asians from this collection are included in POPRES. While extensive cardiovascular-related phenotypic data were collected on these participants, the POPRES database only includes non-identifying demographic information: age at collection, self-identified race/ethnicity, and country of birth.

CoLaus, Lausanne, Switzerland This is a population-based study of European subjects drawn from Lausanne Switzerland, through the CHUV University Hospital [Firmann et al., 2008]. 2,809 subjects from this collection were included in POPRES. While extensive phenotypic data were collected on these participants, the POPRES database only includes non-identifying demographic information: age at collection, self-identified race/ethnicity, native language, country of birth, and parental and grandparental countries of birth.

Duke Healthy Volunteers, NC Healthy volunteers were recruited from the Duke and North Carolina State University campuses. Volunteers were to be aged between 18 and 90 years of age and have no known cognitive impairments. All races and ethnicities were included. 586 subjects from this collection were included in POPRES. Only non-identifying personal demographic information was made available, including and limited to age at collection, self-identified race/ethnicity, and sex.

Informed consent and ethical approval All participants in the component studies that contributed to POPRES provided written informed consent for the use of their DNA in genetic studies. The informed consent form was different for each study, some providing more explicit descriptions of the variety of ways that genotype data derived from the sample may be used than others. Informed consents are available through the dbGaP submission. The informed consents of the Healthy Caucasian Cohort collections were the most extensive. Given the anonymized nature of the collection, these samples were included in POPRES without need for further ethical review. Specific ethical review board approval for the controlled release of de-identified genotype data was sought for the Healthy Taiwanese Cohort, Healthy Japanese Cohort, Healthy Mexican Cohort, CoLaus, and Duke collections. All were granted, with the exception of the Healthy Taiwanese Cohort, which will not be publicly released. The nature of the original consent and ethical review board approval for the LOLIPOP collection was sufficient for the current usage.

Genotyping Genotyping was performed on the Affymetrix (Mountain View, CA) GeneChip 500K Array Set using the published protocol for 96-well plate format. Samples were genotyped in nine batches over a period of 19 months (Table 1.1 with a two to three percent sample duplicate rate to help assess genotype data quality. The CoLaus and LOLIPOP collections were genotyped in multiple batches. All other collections were typed within a single batch. Batch information for each subject is available with the genotype data.

The dynamic model genotype calling algorithm (DM) uses perfect match and mismatch probe intensities to call genotypes for individual arrays. DM was used to measure raw experiment quality. Individual arrays that failed to

Table 1.1: Summary of the studies included in the POPRES study.

Region	Africa	East Asia		South Asia	Latin America
Study	UCSF Af. Am.	Japanese	Taiwanese	LOLIPOP	Mexican
Collection Site	United States	Sydney, Australia	Taiwan	London, England	Guadalajara
Collection Type	Healthy	Healthy	Healthy	Population	Healthy
Sample Size	436	106	174	431	205
500K, Initial QC	346	73	109	360	149
500K, Final QC	346	73	108	359	112
Genotyping batch[a]	9	1	1	7	1
Age (min / med / max)	18 / 45 / 81	> 20	≥ 20	35 / 50 / 74	≥18
Sex (F:M)	279:157	62:44	84:90	121:310	93:112
500K, Initial QC	223:123	44:29	48:61	103:257	69:80
500K, Final QC	223:123	44:29	47:61	103:256	46:66
Call Rate (per SNP)					
Median	0.99	0.99	0.98	0.99	0.98
95th percentile	0.94	0.85	0.87	0.91	0.87

Region	Europe					Mix
Study	USA	Canadian	Australian	LOLIPOP	CoLaus	Duke
Collection Site	North Carolina	Ottawa	Adelaide	London, England	Lausanne, Switzerland	North Carolina
Collection Type	Healthy	Healthy	Healthy	Population	Population	Healthy
Sample Size	27	105	69	938	2809	586
500K, Initial QC	27	105	69	598	2509	490
500K, Final QC	27	105	69	481	2507	-
Genotyping batch2	2	2	3, 7	4, 5, 6	8	
Age (min / med / max)	≥18	≥18	≥18	23 / 54 / 75	35 / 52 / 75	18 / 22 / 79
Sex (F:M)	18:9	63:42	47:22	213:724[b]	1508:1301	331:255
500K, Initial QC	18:9	63:42	47:22	184:414	1350:1159	285:205
500K, Final QC	18:9	63:42	47:22	180:301	1348:1159	-
Call Rate (per SNP)						
Median	1.00	0.98	0.99	0.99	0.99	-
95th percentile	0.85	0.89	0.87	0.92	0.93	-

[a]Batch defined by month that genotyping completed: 1 - Nov. 2005, 2 - Mar. 2006, 3 - Aug. 2006, 4 - Sep. 2006, 5 - Nov. 2006, 6 - Dec. 2006, 7 - Jan. 2007, 8 - Mar. 2007, 9 - May 2007

[b]One subject missing sex information and failed genotyping (i.e. sex could not be inferred)

achieve a 90% DM call rate (at P = 0.26) were generally reattempted in genotyping by re-hybridization. Duplicate concordance for the StyI arrays was distinctly lower than that for the NspI arrays on four plates in batch 7 genotyping of the LOLIPOP collection. The samples on these four plates were re-genotyped on the StyI array using fresh DNA aliquots and performing the Affymetrix protocol in its entirety.

A series of identity checks was performed. Samples were removed if reported gender was inconsistent with X-linked genotypes. Samples with no reported gender were left in the dataset, and their gender was inferred from the genetic data. In addition to the 500K genotyping, a subset of eighty-eight snps were typed with the Single Base Chain Extension (SBCE) as-

say [Chen et al., 2000] for all subjects (43 on NspI, 45 on StyI). The SBCE genotypes were compared with those called by DM on the 500K SNP panel. Samples less than 90% concordant between the SBCE data and the Affymetrix 500K SNP panel data on a single array were removed from the dataset.

Final genotype calling was performed using the Bayesian Robust Linear Model with Mahalanobis distance classifier algorithm (BRLMM). Only arrays passing an 85% DM call rate threshold were input into BRLMM. BRLMM is a clustering algorithm that requires batches of arrays to make calls. Arrays were batched together for BRLMM by plate, with a minimum batch size of fifty. Affymetrix Power Tools v1.4 was used to run BRLMM, with the maximum score threshold set to 0.3. Defaults were used for all other parameters. Any inconsistent genotypes for duplicated samples were removed. Samples were considered successfully genotyped if they passed identity checks and achieved a minimum 95% BRLMM call rate on both arrays after removal of inconsistent genotypes.

There are 500,566 unique markers included in the genotyping array. A set of 3,247 markers identified as mapping to multiple sites on the genome were excluded, leaving 497,625 for subsequent analysis.

Quality Control Genome-wide genotyping with an Affymetrix 500K SNP panel was attempted for all subjects over an 18 month period of time. Two rounds of initial quality control were performed. The first included standard checks. Only subjects with call rates greater than 95% for both NspI and StyI chips and confirmed genotype-sex concordance were retained. Relatedness among subjects was evaluated on the basis of identity-by-descent estimates. This identified 48 closely related subjects, primarily from the Mexican cohort,

that were subsequently excluded. For the LOLIPOP collection, it was determined post-genotyping that some subjects received for the POPRES initiative were not a random sample of the larger LOLIPOP collection. Rather, it consisted of subjects that had been collected early in the project, which had an initial focus on recruiting cardiovascular disease-related patients. A subset of subjects were subsequently selected for inclusion in POPRES with a 6% coronary heart disease (CHD) rate that brought CHD-related endpoints in the dataset in line with LOLIPOP overall. This resulted in the removal of 125 subjects. Preliminary principal component analysis (PCA, see below) within Europeans identified 111 subjects from the European LOLIPOP sample on two genotyping plates strongly correlated with scores on the second component, suggesting a problem with genotype data quality. These subjects were excluded. Two additional subjects were excluded because they had highly negative inbreeding F scores, which were calculated using PLINK [Purcell et al., 2007]. The F scores were twice the magnitude of all other samples, indicating potential contamination. A total of 4,835 subjects (82%) passed this first round of checks. The second round of quality control included further PCA to identify subjects with data quality concerns or misreported genetic ancestry. 4,187 subjects (72%) passed the second round of checks. We note that the Duke data were not available during these further quality control measures and are not included in subsequent analyses. However, the collection is described herein and the genotype data are available with the other POPRES data.

With the set of subjects that passed both initial rounds of quality control, we carried out a series of more stringent quality control steps in an effort to further reduce the likelihood of genotyping errors that could negatively influence genetic studies using these data. First, to overcome concerns that the small batch

sizes used to cluster and call genotypes in the original data set could bias the results (e.g. reference [WTCCC, 2007]), a high performance computing system was used to apply BRLMM to the entire set of files, including data from sample duplicates, for the NspI and StyI chips separately. We refer to the genotypes generated by this combined calling strategy as pooled genotypes and those produced in small groups of samples as batched genotypes. The quality of the pooled versus batched genotype calls were assessed by comparing the sample duplicate concordance and call rates of each (Figure A.1). We found that with the BRLMM default quality threshold of 0.3, the batched genotypes resulted in higher duplicate concordance than the pooled calls (99.66% versus 99.56%) as well as higher call rates (97.66% versus 95.12%). For this reason, we relied on the batched calls for all reported analyses.

We then evaluated the influence of the BRLMM quality threshold on duplicate concordance and its relationship to genotype call rate (Figure A.1). As expected, duplicate concordance increased and call rate decreased as the quality threshold decreased from 0.5 toward zero. Based on the improvement in heterozygote concordance observe (0.98 to 0.99) by decreasing the quality threshold from the initial value of 0.3 to 0.2 with only a modest corresponding decrease in call rates (0.96 to 0.93), we selected the 0.2 threshold for this more restricted data set.

We then excluded 54,191 (10.8%) that had 3 or more discrepancies between the batched and pooled calls or that exhibited a batch call rate below 90%. The pruned SNPs showed lower average duplicate-chip concordance rates (96.6% versus 99.8%) and higher levels of Hardy-Weinberg disequilibrium (20% versus 5% of SNPs with heterozygosity levels above the p¡0.001 threshold). The

remaining SNPs have an average call rate of 97.7% and an analysis of individuals for which duplicate chips were run shows a concordance rate of 99.8%. Our selection of a 90% threshold contrasts with the 95% call rate applied in most other studies using the Affymetrix 500K panel. However, because we use a more stringent confidence score (0.2 versus the BRLMM default value of 0.5) we achieve higher genotype quality (duplicate concordance) with lower call rates (see Figure A.1).

Principal Component Analysis Principal component analysis was conducted using the smartpca software [Patterson et al., 2006] using default settings with no outlier removal. Analysis was carried out following the removal of some apparently related individuals (high identity-by-descent estimates), and individuals identified as outliers in preliminary PCA runs based on regional subsets of the data (e.g. Europe, East Asia, etc). Furthermore, due to the large overrepresentation of UK and Swiss individuals, we randomly selected a subset of 200 UK and 125 French-speaking Swiss subjects. This resulted in a sample of 3,082 POPRES subjects. As a reference, and to provide data from Africans in the analysis, we included genotype data (release 23) on the same subset of SNPs from 207 unrelated subjects from the four core HapMap samples [Altshuler et al., 2005]: Yorubans from Ibadan, Nigeria, Japanese from the Tokyo area, Chinese from Beijing, and CEPH Europeans from Utah. To reduce the linkage disequilibrium between markers, we first used the PLINK software to remove all markers with genotypic r2 greater than 0.8, calculated in sliding windows 50 SNPs wide, shifted and recalculated every five SNPs. This process reduced the number of SNPs analyzed to 286,930.

Previous studies have shown that regions with structural variation such as

inversions can strongly influence PCA results [WTCCC, 2007, Tian et al., 2008]. We found from previous work [Novembre et al., 2008] that plots showing the per-SNP correlation between individual genotype scores (0, 1, or 2) and individual PC coordinates are a useful diagnostic for identifying PCs that might be influenced by long-range LD regions. For instance, in the initial analysis of European samples [Novembre et al., 2008], known inversions on Chromosome 8p23 and 17q21 appear as peaks in the correlation plots for some of the lower PCs (e.g. PC 3). (Alternatively, we could have plotted the absolute values of or the square of SNP loadings from the PCA, but here we used the correlation-based approach because much of this work was done before the release of recent versions of smartpca that provide the SNP loadings). The only strong peaks in the correlation plots within the top seven PCs were for the approximately north-to-south European principal component, which exhibited two large peaks, with p-values of association as extreme as 10^{-40} to 10^{-100}. One of these peaks located at 134.6 - 137.6 Mb on chromosome 2 centered on the LCT gene (136.4 - 136.5 Mb). The other peak on chromosome 6 at 29.1 - 32.8 Mb contained the MHC complex, including the HLA-A, -B, -C, -DR, and -DQ genes.

To assess whether such aberrant regions might influence the PCA results and obscure genome-wide patterns, we performed a second PCA analysis where we first removed SNPs from regions surrounding putative peaks of correlation. Though none of the other first seven PCs, aside from PC 5, showed a strong peak of correlated markers, we conservatively removed all SNPs within 2Mb of a marker highly correlated with any of the first ten principal components. We defined the threshold for calling highly correlated SNPs as being within the top 0.2% of r^2 values for correlations of markers against the given principal component. This process excluded over half of the markers (including the lactase

and MHC regions mentioned above), leaving 226,211 SNPs for the subsequent analysis, and resulted in a final set of 143,893 SNPs after excluding markers using the procedure based on the sliding-window-based pruning step described above. Using this more stringent set of SNPs, we reran PCA on the same set of individuals and found that, aside from negation of the eigenvectors, the PCs revealed the same structure as using the full set of markers, and the first seven principal components had a correlation greater than 0.98 between the two runs. This suggests that the initial PCA was capturing genome-wide patterns of variation rather than patterns localized to specific sets of markers, and the peaks of correlation observed were simply particular sets of markers that happened to be correlated with the population structure (such as in the case of the lactase gene with PC 5, the roughly north-to-south European PC). Although the results were similar between the two runs, we present the results from the second of the two PCA runs.

Case-cohort Matching and Genome-wide Analysis We performed four different methods of case-cohort matching to assess their impact on type I error rates in an example motivated by the search for major genetic risk factors for adverse drug reactions. Twenty-two HIV-positive patients of European origin with clinically diagnosed abacavir-associated hypersensitivity reaction were genotyped with the Affymetrix 500K SNP panel as previously described [Nelson et al., 2008]. One case was dropped due to very low genotyping efficiency (<85%). Ten cohort individuals were matched to each case by four methods: 1) continental origin, selecting Europeans from the United Kingdom, 2) country of sampling or country of birth (if available), 3) minimizing pairwise

identity by state (IBS) distance, and 4) minimizing pairwise distance among selected principal components.

Continent of origin matching was carried out using POPRES subjects of self-identified European origin who were collected in, or reported to have ancestry from, England or the United Kingdom. Country matching was carried out by selecting sex-matched cohorts from the same country of origin as the cases (Table A.1). When there were excess numbers of cohort individuals available, ten were randomly selected for each case. Cohorts from adjoining countries were selected when there were insufficient numbers of cohorts available from the case countries. IBS matching was carried out by estimating the pairwise IBS distance from each case to each POPRES subject that satisfied the QC criteria described above. IBS estimation was carried out with PLINK v1.01 [Purcell et al., 2007], excluding 58,089 SNPs found within genomic regions highly correlated with the scores from the top four PCs in a European-only analysis (as described above), 61,275 SNPs missing more than 5% of genotypes, and 96,880 SNPs with minor allele frequencies less than 5%. For each case, the ten POPRES subjects with the shortest IBS distance to the case were selected as controls. PCA matching was carried out using PCA scores. PCA, excluding 58,089 SNPs described above, was carried out on the combined cases and subset of POPRES defined as European origin, with analysis limited to 200 subjects per country and principal component scores assigned to all eligible controls. Inspection of the resulting eigenvalues led to the selection of the first four components for genetic matching. Prior to matching, eigenscores were rescaled to reflect their relative importance by multiplying each eigenscore by the square root of the corresponding eigenvalue. Pairwise Euclidean distances were then estimated between each case and all POPRES subjects. Ten cohort individuals were selected for each

case, randomly selecting cohorts within the 2.5th percentile of the multivariate distance distribution, with care not to allow the reuse of cohorts among cases.

For each of the four selections of cohorts, genome-wide association analysis was carried out using Fishers exact test, as described previously [Nelson et al., 2008]. SNPs were excluded from analysis if they were missing mapping position, had genotyping efficiency less than 90%, had minor allele frequency less than 1%, or had deviations of genotype frequencies from Hardy-Weinberg expectations that were highly significant (p-value $< 10^{-7}$) in cohorts. We also excluded 26 SNPs identified in a previous study to have highly erroneous genotype calls within the 21 cases [Nelson et al., 2008]. Comparisons across analyses were carried out on a final set of 393,699 SNPs that passed the QC in all four case-cohort samples.

Public Data Availability The subject-level data described in this study is available via the dbGaP archive sponsored by the National Center for Biotechnology Information (see Web Resources) pending acceptance of a standard Data Use Certification and endorsement by the requesting investigators institution. Data include the demographic variables listed in the following section, PCA scores, and genotype data described herein.

1.4 Results

Sample and Data Overview The POPRES study includes DNA samples from 5,886 subjects derived from ten constituent collections (Table 1.1; the LOLIPOP study is divided between subjects of Indian Asian and European origin). Based

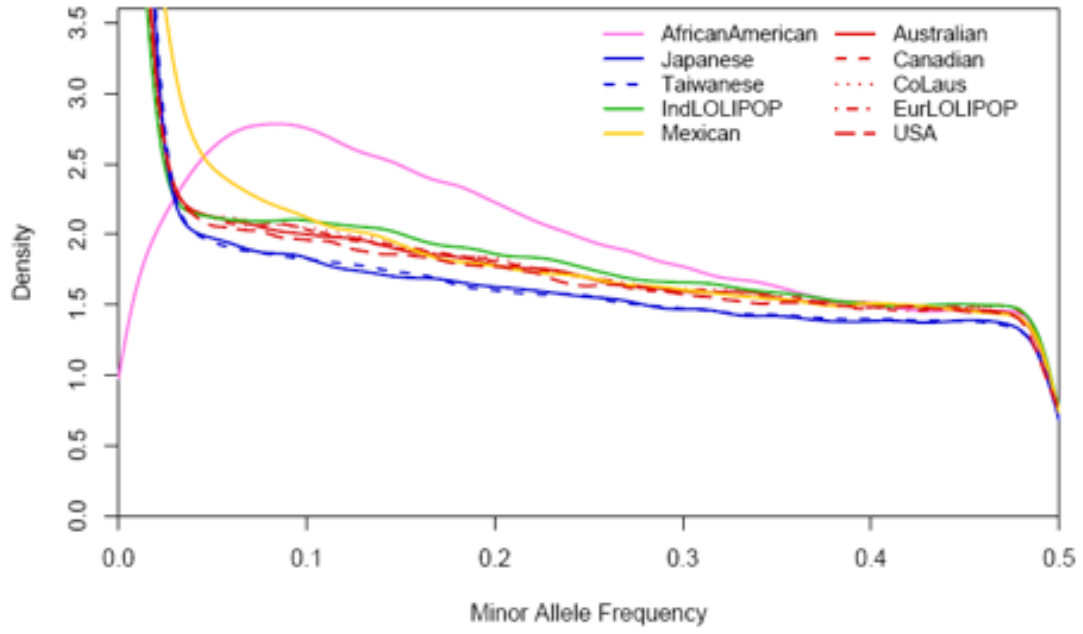

Figure 1.1: Distribution of minor allele frequency by collection. Colors and line types for the densities of each collection are shown within the figure.

on the inclusion criteria and recruiting methods, these collections are broadly described as either population samples or healthy subjects (see Methods for collection-specific details). Basic demographic data available for all subjects includes sex, country of collection, and self-described racial background. Additional information available for some collections includes age at collection, state or city of collection, country of birth, parental country birthplaces, grandparental country birthplaces, and native language. Complete demographic summaries of each collection are provided in the Supplementary Results and subject-level details are available via controlled access in a public repository (see Web Resources). All participants were at least 18 years of age at time of recruitment. The sex ratio varies widely among studies.

The distribution of minor allele frequencies by collection is presented in Figure 1.1. The frequency distributions are very consistent among the five European collections as well as between the two East Asian collections. However, the distributions differ substantially among the five major geographic regions shown. East Asia shows the highest proportion of low frequency SNPs (22% of SNPs less than 0.01 frequency), followed by Europe (15%), South Asia (13%), Mexico (10%), and lastly African American (1.9%). These frequency distributions differ markedly from those observed in the resequenced ENCODE regions of the HapMap project [Altshuler et al., 2005], wherein Europeans showed an increase in low frequency SNPs compared to East Asians and levels comparable to Africans. These differences reflect the biased nature of the SNPs included on the genotyping array [Clark et al., 2005].

The distribution within African Americans is most distinct. There are a large proportion of SNPs with frequencies between 0.05 and 0.2, which is consistent with the African HapMap ENCODE and Affymetrix 500K SNP data (Figure A.2). However, the African Americans have a very small proportion of low frequency and monomorphic SNPs compared to the other continental groups and compared to HapMap Africans (Figure A.2). This does not reflect the underlying SNP frequency distribution in African Americans [Lohmueller et al., 2008], but rather the influence of African and European admixture of African Americans with the SNPs in this panel. Although 15% of these SNPs have minor allele frequencies less than 0.01 in Europeans and 11% in YRI, only 1.6% of them have minor allele frequencies less than 0.01 in both. This smaller proportion of low frequency SNPs suggests that this panel would be more informative for studies in African Americans, compared to Africans.

Analysis of population structure We performed a principal component analysis (PCA) on the genotype data to investigate the main axes of variation present in this sample. PCA makes inferences solely on the genotype data without inclusion of any other information; hence the analysis results reflect the clustering within those data. The results of the PCA with the POPRES and HapMap data combined exhibit the anticipated structure of first clustering continents and next regions within continents (Table 1.2, Figures 1.2 and A.3). As expected, the first principal component (PC 1) distinguishes Africans from non-Africans. The next three principal components also characterize continental regions: PC 2 distinguishes East Asians from Africans and Europeans, with South Asians and Mexicans at intermediate values; PC 3 distinguishes South Asians from East Asians; and PC 4 distinguishes Mexicans from non-Mexicans.

The subsequent principal components mark within-continent variation. PC 5 reveals a north to south cline within Europeans (Figure 1.3) consistent with existing studies of European substructure [Tian et al., 2008, Novembre et al., 2008, Bauchet et al., 2007]. The majority of Europeans sampled from North America and Australia are most similar to northern Europeans, with modest numbers of outlier observations. The CEU sample had the highest median scores on this component, followed by Australia and USA (collected in North Carolina), then by Canada, having a median more similar to central than to northern Europe. PC 6 distinguishes the African Americans from the HapMap Africans. Interpreting the asymmetrical distributions of the Africans and African Americans along the European north-south cline in Figure 1.2C suggests that the Africans are slightly more similar to southern Europeans, whereas the African Americans lie slightly shifted to the right, and on average appear more like northern Europeans on this principal component. This may be partially due to northern

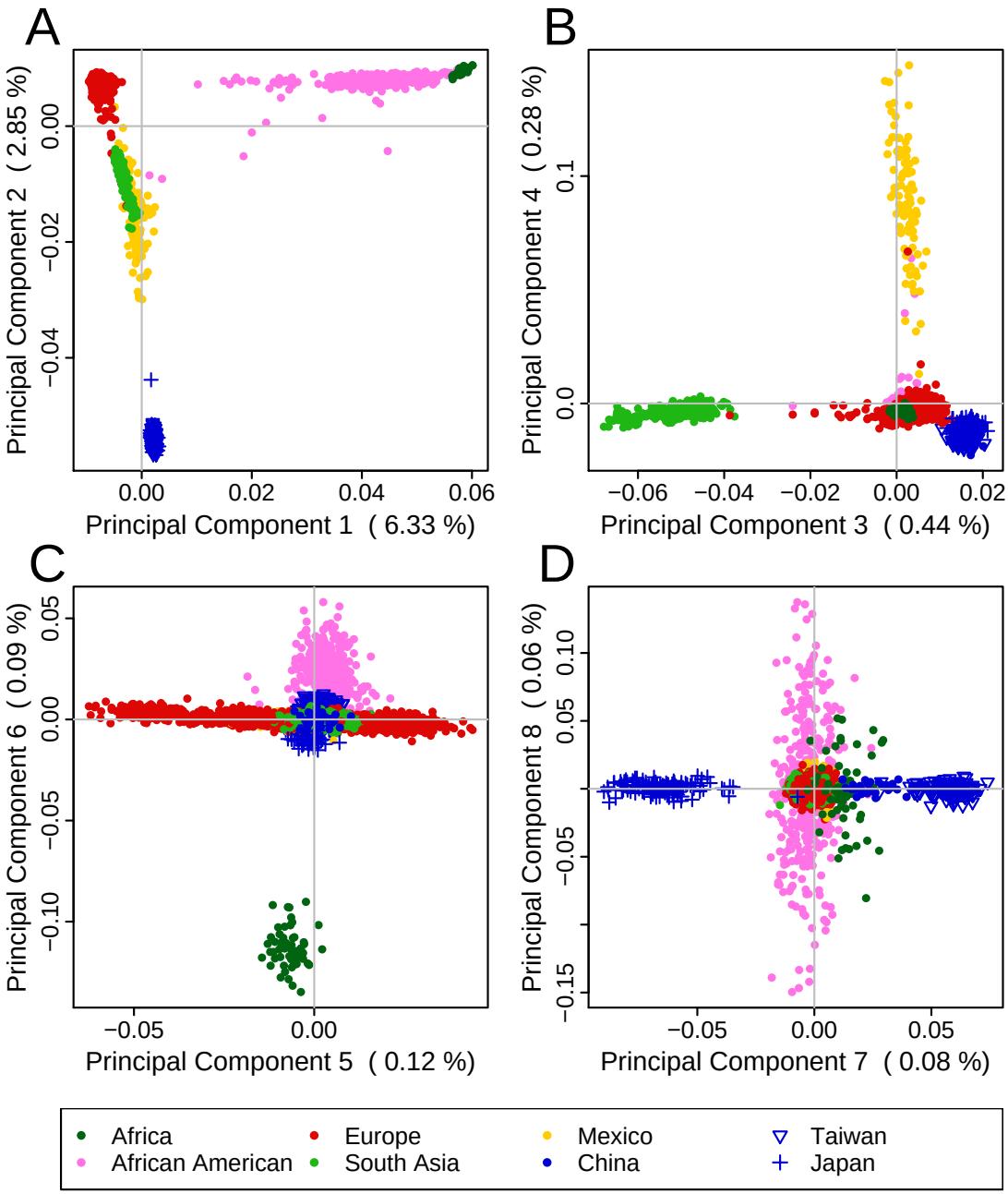

Figure 1.2: Population genetic structure illustrated through scatter plots of consecutive principal components. Subject scores are colored by continental/ethnic origin (see legend). East Asian populations are indicated by varying point types. Percent of variation explained by each component given in parentheses on each axis label.

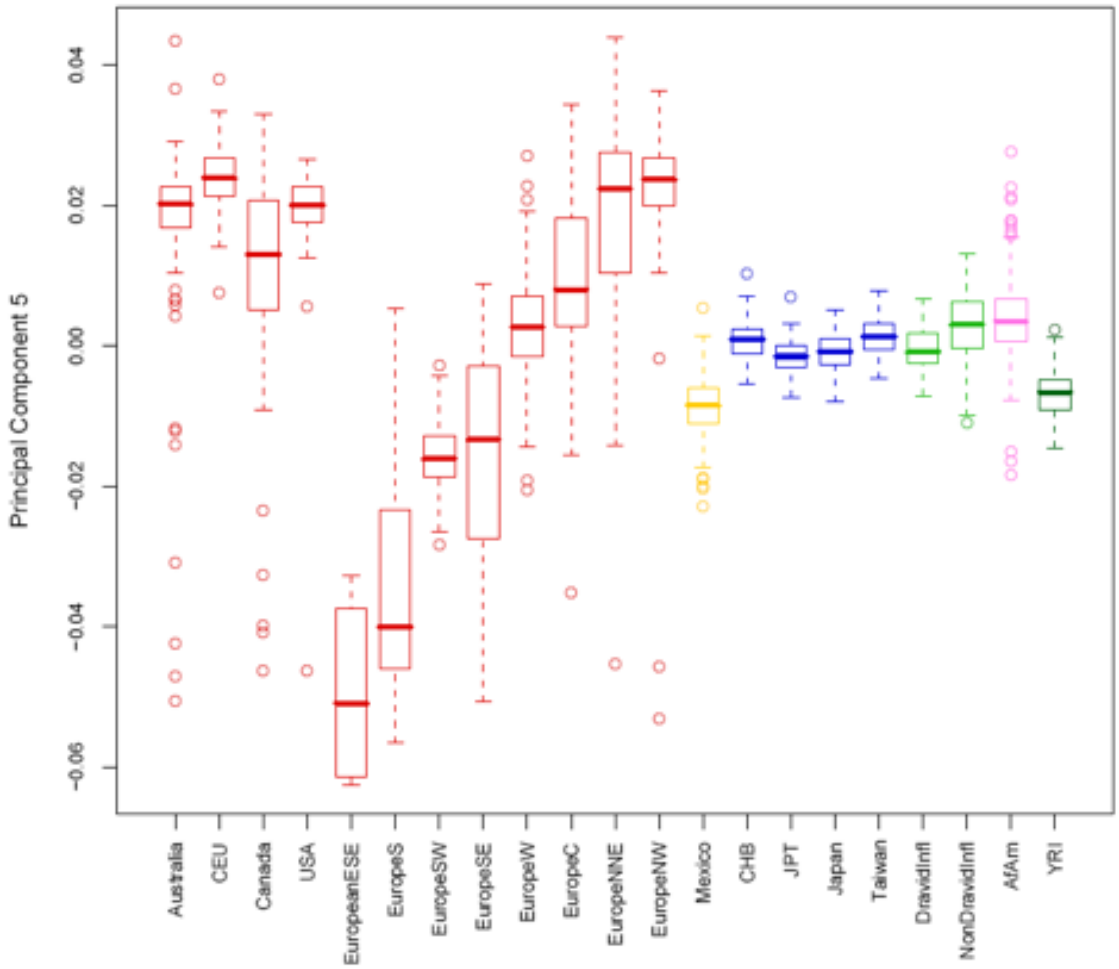

Figure 1.3: Distribution of subject-level principal component 5 scores by reported ancestry. Each box and whisker indicates the median (heavy line), interquartile range (IQR, box), and minimum/maximum observations (whiskers). Whiskers are truncated at the last observation within 1.5 times the IQR from the edge of the box, with outliers shown individually.

European admixture in African Americans. However, caution should be used in this interpretation, since the Africans and African Americans are slightly more similar to their respective subpopulations of Europeans only on genotypes that distinguish southern from northern Europeans, and this similarity is not necessarily true of overall genotype relatedness.

Principal component 7 (Figure 1.2D) separates Japan (left), from the HapMap

Table 1.2: Summary of regions distinguished by the top principal components.

PC	% Var Explained	Subpopulations Distinguished	Influential Genomic Regions
1	6.33	Africa / African Americans	-
2	2.85	East Asia	-
3	0.44	South Asia	-
4	0.28	Mexico	-
5	0.12	N Europe vs. S Europe	LCT, MHC
6	0.09	Africans vs. African Americans	-
7	0.08	Japan vs. Taiwan vs. China	-

CHB on mainland China (center right) from Taiwan (far right). Note that the Africans, unlike African Americans or other continents, appear more similar to the Chinese than Japanese on the PC that distinguishes East Asian substructure. We omit showing further results, as PC 8 and subsequent PCs show substructure within Africans and African Americans not corresponding to any known geographic or population structure among individuals. The first two PCs explain a total of 9.2% of the genetic variation within this sample. The remaining five PCs, though clearly informative, only explain an additional 1.0% combined.

Case-cohort matching One of the primary motives in the development of the POPRES resource was to provide a source of pre-genotyped population samples that could be drawn on as needed as a comparator (i.e. cohort) group for association studies of adverse drug reactions (ADRs). The rationale for this approach and its implications on statistical power for ADR genetics research has been considered elsewhere [Nelson et al., 2008]. In that previous work, we argued that use of population cohort required that they be matched appropriately to the cases. Given such a resource, there are multiple ways in which cases and cohorts could be matched. Here, we extend our previous work with 21 clinically diagnosed abacavir-associated hypersensitivity reaction (ABC HSR) cases [Nelson et al., 2008] by comparing four strategies for matching them to

these POPRES cohorts: 1) matched by continental origin by selecting northern Europeans from the United Kingdom, 2) matched by reported country or region of birth, 3) minimizing pairwise identity-by-state (IBS) distances between cases and cohorts (Figure A.4), and 4) minimizing distances between cases and cohorts based on multivariate PCA scores (Figures A.5 and A.6). For each method, cohorts were matched to this small sample of cases in a 10:1 ratio.

The results of each genome-wide association analysis, using cohorts selected as described above, are summarized in Figure A.7. All four methods identify the known MHC region (tagging HLA-B*5701) among the top 20 associated SNPs, with PCA matching yielding the lowest p-value and highest rank (p-value = 2.1×10^{-6}, rank = 2), followed by UK (4.2×10^{-6}, 8), country (7.6×10^{-6}, 5), and IBS (2.9×10^{-5}, 16) matching. A comparison of the ranking amongst the top 100 SNPs from each analysis showed that the country and IBS matching methods were the most concordant (ρ = 0.58). The comparison between country and PCA matching were the least concordant (ρ = 0.03). The remaining pairwise comparisons were only modestly correlated ($\rho < 0.15$).

With a single realization of each matching algorithm, it is not possible to assess the impact of the matching on the power to identify the known effect of the HLA-B*5701 allele. However, with nearly 400,000 SNPs for which the null hypothesis of no association is true, we can reasonably assess the effect of each matching algorithm on the type I error rate. The proportion of tests with p-values falling below a range of significance thresholds, shown in Figure 1.4, is very similar amongst the country (genomic control λ = 1.00 for allelic test), IBS (λ = 1.00), and PCA (λ = 1.00) matching methods and falls close to the expected proportion at each level. In contrast, the analysis that only drew from

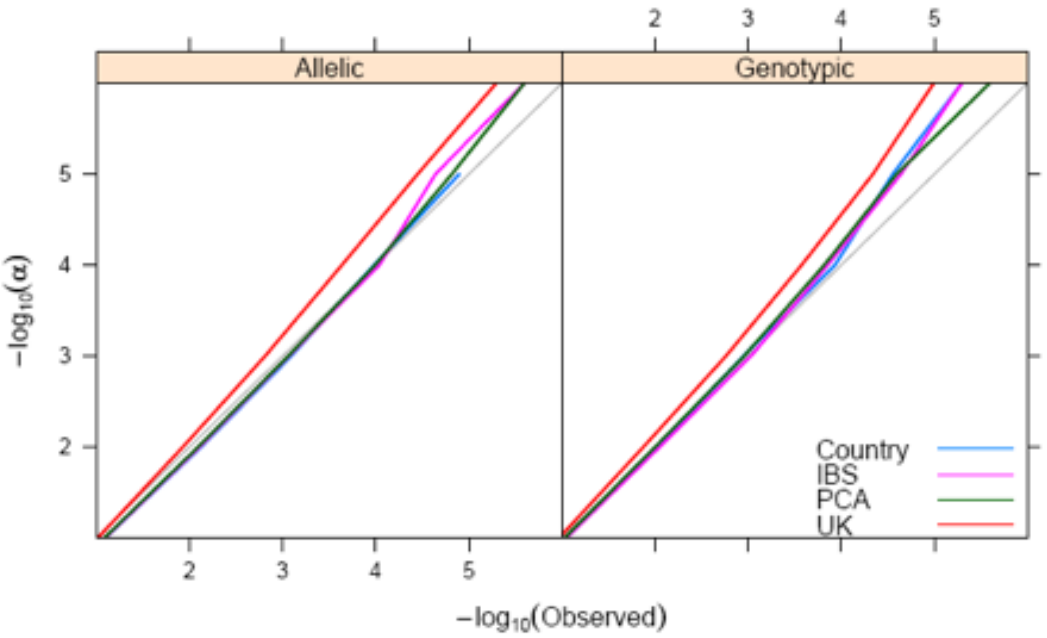

Figure 1.4: Comparison of observed versus expected proportion of associations over a range of significance thresholds (P-P plot). Separate lines are presented for each of the four cohort matching strategies. Results of the allelic exact test are shown on the left and genotypic exact tests on the right. A light gray line corresponds to unity.

population cohorts in the UK ($\lambda = 1.13$) resulted in a significant excess of low p-values at all levels below 0.1, roughly doubling the numbers observed with the other matching methods. While all four cohort matching procedures resulted in relatively low p-values for the known association, the UK cohorts (i.e. matching only by continent) suffered from an increase in the false positive rate, even with this small number of cases. Figure A.7 shows that relatively small p-values are observed across the genome and vary substantially across cohort selections.

1.5 Discussion

We have brought together DNA from nearly 6,000 subjects participating in ten studies with ancestry from five major geographic regions and dozens of countries as a resource for genetics research. Genotype data from a genome-wide panel of 500,000 SNPs attempted on nearly all participant samples were carefully evaluated to yield a set of subjects and markers with high data quality that may be appropriate for a range of applications. These data are freely available for legitimate research purposes through the public dbGaP website.

Principal component analysis (PCA) of these data illustrates the overall data quality, both in terms of genotypes and labels of subject origins. The seven highly informative principal components (PCs) provided a high degree of discrimination among African, East Asian, South Asian, European, and Mexican ancestry. It also illustrated finer differentiation between African versus African American, among Japanese, Han Chinese from Beijing, Han Chinese from Taiwan, and highlighted genetic gradients within African Americans, Mexicans, and Europeans. These results provided ample opportunities to identify subjects with ancestry labels that match their genetic background. Very few subjects demonstrated PC score patterns that deviated noticeably from the majority of their groups. The score information (available via dbGaP) may be used in future applications to re-label subjects or to exclude them from further analyses.

The potential impact of cases and cohorts that are poorly matched for their genetic background on the type I error rates of association studies is well understood (e.g. [Lander and Schork, 1994]). Most studies of unrelated subjects attempt to cohorts for this through careful study design and sam-

pling (e.g. [McCarthy et al., 2008]), statistical correction (e.g. 26), or measuring and correcting sample structure by use of PCA or related methods (e.g. [Price et al., 2006, Yu et al., 2006]). Alternatively, sets of healthy or population controls that have been genotyped for compatible genome-wide panels can be queried for controls that genetically match the genotyped cases [Luca et al., 2008, Hinds et al., 2004], recently illustrated for genome-wide genotype data [Luca et al., 2008]. In the limited application presented here using 21 subjects with abacavir-associated hypersensitivity reaction, we found that matching cohorts to cases based on country of origin, minimizing pairwise IBS distances, and minimizing distances among the top principal components were similarly effective in controlling type I error. The latter two genotype-based methods would clearly be preferred when there is uncertainty about genetic background of the cases or controls, or when the populations sampled are admixed or otherwise genetically heterogeneous. It is important to note that with such a small number of cases included in this example application, there is insufficient power for subtle population or genotype quality-dependent differences between the cases and controls to be detected. An analysis with a larger number of cases and controls could highlight limitations in the sample matching schemes or in the POPRES data that were not readily apparent in this example.

Most studies that include whole-genome genotype data do not have need for external sources of controls for key analyses, and even with 5,000 subjects genotyped, the power of this resource to investigate common disease genetics is limited, particularly for non-European populations. Nevertheless, the data published herein should prove useful for characterizing the genetic background of study participants, particularly for small sample sizes or poorly characterized sample collections. POPRES genotype data may be included with study geno-

type data to conduct analyses of population structure. The Affymetrix 500K SNP panel shares a reasonably large number of SNPs with other popular SNP panels, including Illumina 1M (138,143) and Affymetrix 6.0 (469,874), which have been shown to be generally sufficient for inferring patterns of population structure at several scales [Auton et al., 2009]. The legitimacy of this approach is most obvious for genotype data derived from the Affymetrix 500K and 6.0 SNP panels. However, it should be possible to derive informative subject scores with this approach from the subset of SNPs that overlap with the Illumina panels, though the accuracy of this approach has not be assessed. Beyond the global patterns of variation observed in the analyses included in this report, finer-scale structure may also be investigated in subsets of the POPRES data, such as within Europeans [Novembre et al., 2008].

As described, nearly all of the subjects currently included in POPRES have been genotyped with the Affymetrix 500K SNP panel. The choice to standardize on this panel was largely influenced by the timing of the project. Since the time this project was initiated, genome-wide genotyping panels from multiple vendors have expanded and improved in quality. Although there is no expectation that the entire POPRES collection will be genotyped on another genome-wide panel, selected subsets will be genotyped with newer panels as required to support ongoing research, and much of these data will eventually be deposited to dbGaP. This includes existing data on the Illumina (San Diego, CA) 550K and 1M panels typed on 500 POPRES subjects of European origin. Developments around use of representative patterns of haplotype structure to impute unmeasured genotypes may also be employed with this and similar resources to make the results from the Affymetrix 500K panel compatible with other panels [Marchini et al., 2007, Browning and Browning, 2007, Scott et al., 2007].

In developing this resource, we considered several alternative designs. The first objective is to use this collection as a resource for generating contrast (cohort) groups for pharmacogenetic studies. In the context of studying the occurrence of an ADR , the cohorts would ideally match the cases for disease status, treatment, duration of treatment, age, gender, and any other disease- or ADR-related clinical characteristics so that associated markers can be inferred to be causally related. However, developing a general resource applicable to a diversity of diseases and relevant to a number of drugs (approved or in development) would likely require extremely large samples and be difficult, if not impossible, to ascertain. When the outcome under study is relatively rare (prevalence <10%), as many ADRs are, an alternative to having treatment-matched patients is having patients matched for disease status, but unknown for their propensity for an adverse event given the lack of treatment. Because the outcome is rare, a relatively small percentage of the cohorts would have had the adverse event, if they had been treated. This more feasible design would result in little loss of power to detect even modest genetic effects. Even so, unless the number of relevant diseases is very small and foreseeable, even large collection sizes will be limited once study-specific strata are considered.

With these limitations, we considered that a collection representative of the populations from which the cases were sampled without regard to disease status may be the most feasible design. A population sample design would result in disease frequencies in similar proportions as the population at large. For rare outcomes, the frequency of those genetically predisposed to the outcome of interest would be low, resulting in a small loss of power to identify predisposing factors. In this design the disease status and outcome of interest are likely to be confounded requiring further investigation to disentangle the relevance of each

result.

It is often of interest to estimate the frequencies of alleles associated with a pharmacogenetic response. One can estimate these frequencies in the population of affected individuals (i.e. patients) or in the population at large. While estimates in patients are more representative of the intent to treat population, having an appropriate sample for a large number of diseases is not feasible. Estimating the genetic parameters in the population at large will only be limiting if the genetic variant, or one in linkage disequilibrium with it, plays an important role in both the disease susceptibility as well as in the pharmacogenetic response under investigation. This may be expected to occur when the variations with pharmacogenetic impact are located within the drug target. In such cases, caution should be exercised in the interpretation of results.

The range and value of genetic studies possible using such a resource rests largely on the quality, quantity, and sampling of the data available. The public release of the POPRES resource will have immediate opportunities to impact a variety of studies and contribute to the growing body of data that will further many areas of human genetics research. We support the public access to these data for appropriate research uses and encourage the further development of such resources for the benefit of the scientific community.

Web Resources

Controlled access to demographic and genotype data via dbGaP:

`http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap`

Acknowledgements

We thank Clive E. Bowman, Michael Klotsman, Ann Marie McNeill, David P Yarnall, Ross Haggart, Steve Haneline, Kelley Johansson, Devon Kelly, Devi Smart, Sarah Tate, Jill Ratchford, Mike Lawson at GlaxoSmithKline for their many contributions to the development of the POPRES resource. We thank Arlene Hughes and Bill Spreen at GlaxoSmithKline for their work on abacavir-associated hypersensitivity reaction pharmacogenetics research. We gratefully acknowledge Yolande Barreau, Anne-Lise Bastian, Binasa Ramic, Martine Moranville, Martine Baumer, Marcy Sagette, Jeanne Ecoffey and Sylvie Mermoud for theirs roles in the CoLaus data collection. The CoLaus study was supported by research grants from GlaxoSmithKline and from the Faculty of Biology and Medicine of Lausanne, Switzerland. We thank Anna C. Need for providing access to the Duke healthy cohort collection. We thank Robin Lincoln at UCSF for expert specimen management at UCSF. Recruitment of the UCSF African American samples was funded by grants from the National Institute of Health (RO1 NS046297) and National Multiple Sclerosis Society (RG3060C8).

CHAPTER 2

GLOBAL DISTRIBUTION OF GENOMIC DIVERSITY UNDERSCORES RICH COMPLEX HISTORY OF CONTINENTAL HUMAN POPULATIONS*

*Originally published as: A. Auton, K Bryc, A. R. Boyko, K. E. Lohmueller, J. Novembre, A. Reynolds, A. Indap, M. H. Wright, J. D. Degenhardt, R. N. Gutenkunst, K. S. King, M. R. Nelson, and C. D. Bustamante (2008). *Genome Res*, 19(5):795803

2.1 Abstract

Characterizing patterns of genetic variation within and among human populations is important for understanding human evolutionary history and for careful design of medical genetic studies. Here, we analyze patterns of variation across 443,434 SNPs genotyped in 3,845 individuals from four continental regions. This unique resource allows us to illuminate patterns of diversity in previously under studied populations at the genome-wide scale including Latin America, South Asia, and Southern Europe. Key insights afforded by our analysis include quantifying the degree of admixture in a large collection of individuals from Guadalajara, Mexico; identifying language and geography as key determinants of population structure within India; and elucidating a North-South gradient in haplotype diversity within Europe. We also present a novel method for identifying long-range tracts of homozygosity indicative of recent common ancestry. Application of our approach suggests great variation within and among populations in the extent of homozygosity suggesting both demographic history (such as population bottlenecks) and recent ancestry events (such as consanguinity) play an important role in patterning variation in large modern human populations.

2.2 Introduction

Recent advances in sequencing and genotyping technology have transformed the study of human population genetics [Frazer et al., 2007, Hinds et al., 2005]. Analysis of dense genotype data has greatly expanded our understanding of the role natural selection has played in the recent evolution of our species

[Sabeti et al., 2007, Voight et al., 2006, Williamson et al., 2007], the nature and causes of recombination rate variation [Myers et al., 2006, Coop et al., 2008], and the extent of structural variation within and among human genomes [Redon et al., 2006, Kidd et al., 2008, Jakobsson et al., 2008].

Arguably, some of the most important insights have come from refining our views of human population structure and recent demographic history [Altshuler et al., 2005, Schaffner et al., 2005, Frazer et al., 2007, Keinan et al., 2007, Jakobsson et al., 2008, Li et al., 2008]. For example, the HapMap Project [Altshuler et al., 2005, Frazer et al., 2007] has afforded unprecedented insight into fine-scale patterns of genotype and haplotype variation across more than 3.1 million single nucleotide polymorphisms (SNPs) genotyped in 270 individuals from three major continental populations. Likewise, analysis of samples collected by the Human Genome Diversity Project (HGDP) [Jakobsson et al., 2008, Li et al., 2008] has elucidated patterns of diversity across approximately 650K SNPs genotyped in nearly a 1,000 individuals from 51 populations. One key feature of these projects is that they have focused on comparing geographically discontinuous populations with small to moderate sample sizes per group. They have also specifically excluded individuals of admixed ancestry in many of their analyses.

In this paper, we analyze dense genotype data from 3,845 individuals from the Population Reference Sample (POPRES [Nelson et al., 2008]), with self-identified ancestry from four continental regions (Table B.2). The POPRES is comprised of samples from a number of studies and includes both individuals designated as healthy, and individuals undisclosed disease status [Nelson et al., 2008]. Individuals were generally sampled in urban locations,

and were genotyped on the Affymetrix GeneChip Mapping Array 500K. Depending on the original study for which the samples were collected, further non-genetic data are often available, including self-reported ancestry up to and including grand-parental information, and primary spoken language.

The POPRES study provides a complementary resource to both the HapMap and HGDP datasets, and presents an opportunity to further understand human genetic diversity. In this paper, we have investigated population structure, haplotype diversity and patterns of homozygosity in the POPRES. Some of the key findings we have uncovered using the POPRES data include:

- Consistent with previous studies, we find F_{ST} to be low between human populations. However, we find F_{ST} to be higher than expected on the X-chromosome.
- Evidence of historical South European admixture with the Mexican population. We estimate an average of 32.5% European ancestry in individuals of Mexican origin, with large variation between individuals in admixture proportion.
- Population stratification within South Asia. Specifically, we observed clustering of individuals who speak Dravidian-influenced languages spoken in southern India.
- Higher haplotype diversity and African haplotype sharing in South and South-West Europe compared to South-East Europe, consistent with gene flow across the Mediterranean.
- Evidence for runs of homozygosity (ROHs) in almost all individuals examined, with striking variation in ROHs among individuals and between populations.

Together these analyses suggest the growing utility of large diverse samples of world-wide human populations, such as the POPRES collection.

2.3 Results

Population Structure Consistent with all previous studies of human genetic variation, we find that the vast majority of common genetic variation is shared across major continental populations. Specifically, we observed a low degree of population differentiation, as measured by Wright's fixation index, of F_{ST} = 5.2% across autosomal SNPs for the four main continental groupings of East Asia, South Asia, Europe, and Mexico. Interestingly, we observed a significantly higher degree of divergence in allele frequency across X chromosome SNPs where we estimate F_{ST} to be 9.7%. This value is about 40% higher than the expected value of 6.8% derived from a island model and accounting for the 4:3 ratio of autosomes to sex chromosome. The higher degree of population divergence at X chromosome SNPs relative to autosomes suggests a smaller effective population size of the X than that predicted from Mendelian genetics, but could also be explained by region-specific selection, sex-biased migration or other demographic forces.

In order to quantify patterns of population structure and admixture, we utilized *STRUCTURE*, a commonly used Bayesian clustering method. Due to computational limitations of the algorithm, we applied *STRUCTURE* to a subset of the data [Pritchard et al., 2000]. Specifically, we randomly selected 6,567 SNPs with MAF > 0.2 and spacing of at least 400kb (See Methods). For comparison and further validation of the POPRES data, we also included the four HapMap

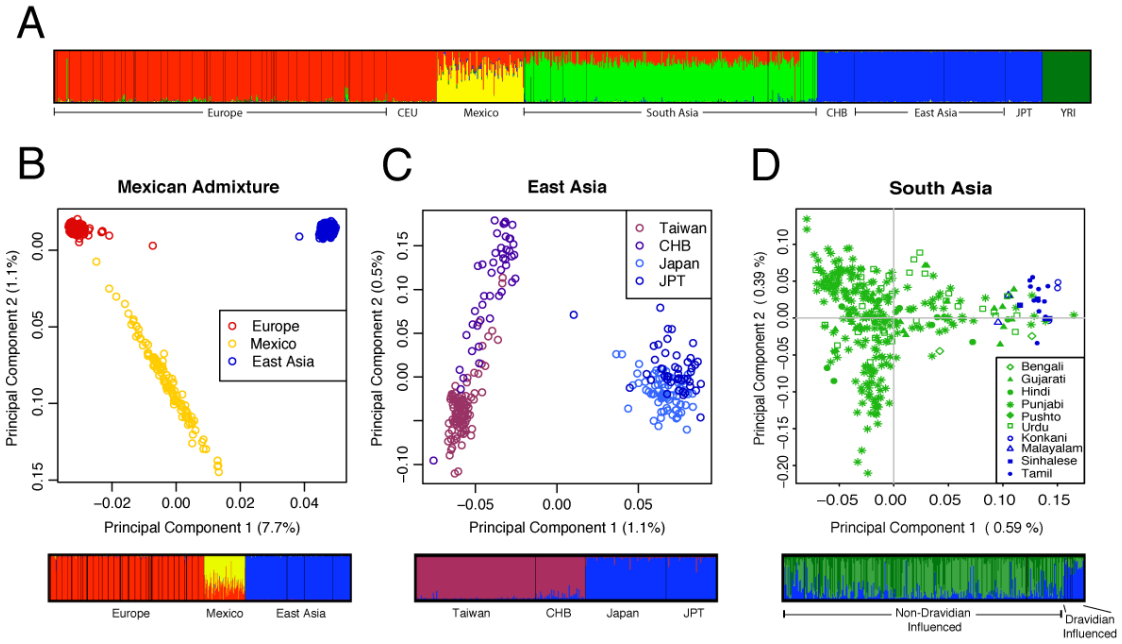

Figure 2.1: Global and regional patterns of population structure. (A) *STRUCTURE* analysis with $K = 5$ for the POPRES populations combined with the HapMap populations. (B, C and D) For each region, the first two principle components are shown, with the proportion of variance explained by each component shown in brackets. Results from *STRUCTURE* are shown below the PCA results, with K=2 for East Asia, and K=3 for South Asia and Mexico. HapMap samples have been included in the East Asia analysis for comparison. In South Asia, speakers Dravidian Influenced languages are shown in blue, whereas Non-Dravidian languages are shown in green.

(release 23) populations in this analysis using the same SNP subset. Setting the number of clusters (K) to five revealed structure largely corresponding to continental regions (Figure 2.1A). Interestingly, all Mexican and many South Asian individuals showed a proportion of the genome clustering with European individuals. In the case of individuals from Mexico, the European component most likely reflects recent admixture, whereas the smaller European component in South Asia perhaps represents the recent common ancestry of the two populations [Patterson et al., 2006]. This is apparent in a Principal Component Analysis (PCA) of the Mexican population combined with the European samples (Figure 2.2A) with Mexican individuals forming an elongated cluster extending

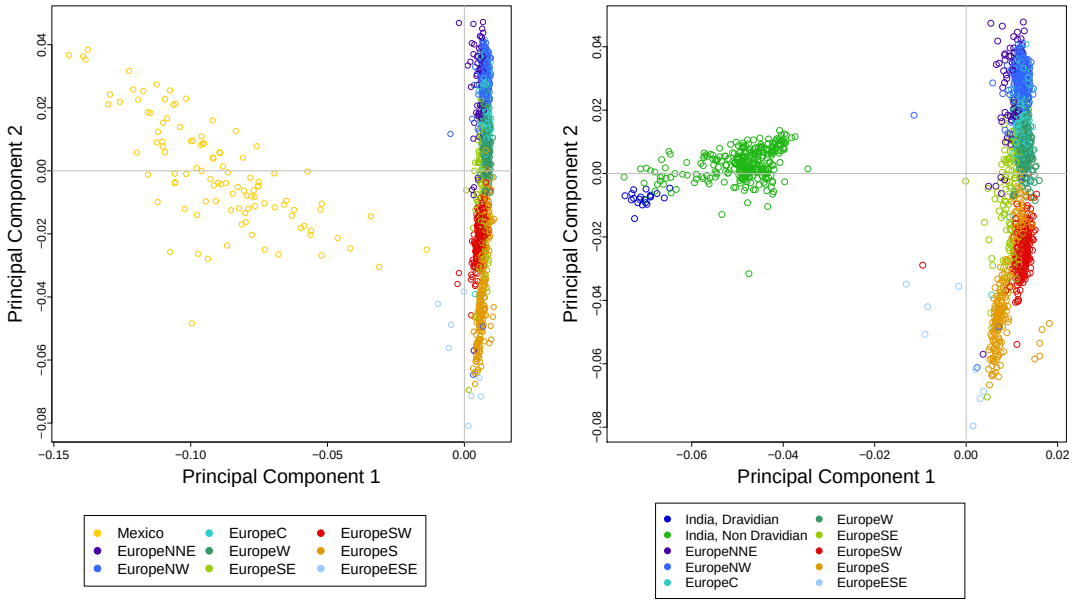

Figure 2.2: Principal Component Analysis of Europe and Mexico (left), and Europe and South Asia (right). Each point represents an individual, and is colored by the assigned population group.

from South / South West Europe. Conversely, in similar analysis, South Asians form a tighter cluster that exhibits no preference for any one region of Europe (Figure 2.2B). However, weak structuring by spoken language group is visible within the South Asian cluster, which is consistent with geographic structure (see Appendix B).

To investigate the level of admixture in the Mexican population, we combined the Mexican samples with a sample of European and East Asian populations. Using *STRUCTURE* with $K = 3$ we estimated an average of 32.5% European ancestry in Mexican individuals (± 3.3% 95% C.I.; see Figure 2.1B), which is lower than some previous estimates based on microsatellite or 'ancestry informative' markers [Wang et al., 2008, Price et al., 2007, Salari et al., 2005, Tian et al., 2007].

Analysis of the East and South Asian populations reveals structuring at a

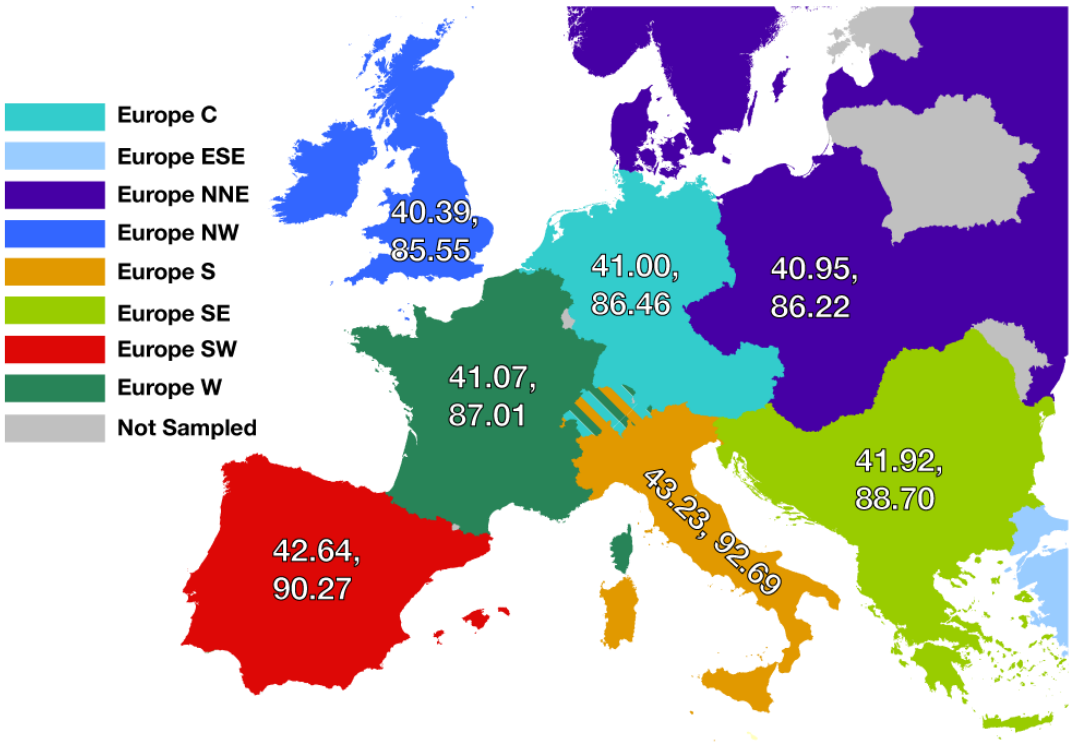

Figure 2.3: Haplotype Diversity within Europe. Geographic regions are color coded. Individuals from Switzerland (striped region) were grouped into adjoining regions on the basis of spoken language. Two numbers are shown within each region, with the first representing H_{10} and the second representing H_{25}.

subcontinental level. In East Asia (Figure 2.1C), we observe clear separation of the Japanese populations from the Taiwanese and HapMapCHB populations, with weaker separation of the Taiwanese from the CHB. In South Asia (Figure 2.1D), we observe weak clustering by spoken language. Assuming language to be a reasonable proxy for geographic location, the observed structure is therefore correlated with geographic spacing as seen in studies using fewer of markers [The Indian Genome Variation Consortium, 2008, Kashyap et al., 2006, Basu et al., 2003]. Furthermore, of the two South Asian populations, the Dravidian Influenced group is slightly more diverged from the other continental populations (Table B.3), suggesting stronger genetic isolation of this population.

To understand how representative the POPRES samples are of human diversity, we combined the POPRES dataset with the HGDP dataset [Jakobsson et al., 2008]. These two studies used very different sampling strategies, with the POPRES mainly sampling individuals in urban locations, and the HGDP focusing on more isolated populations. However, analysis of the combined dataset revealed global patterns of population structure consistent with those previously observed, with the POPRES samples clustering with the corresponding HGDP populations (Figure B.4). It is perhaps worth noting that the HGDP populations form tighter clusters than the POPRES populations, which is to be expected given the disparate sampling schemes of the two studies.

Patterns of Haplotype Diversity Patterns of LD among SNPs provide important information regarding human evolutionary history. As a summary of LD within populations, we considered the average haplotype diversity in each population. We chose to summarize haplotype diversity for each population by the average number of distinct haplotypes in 0.5 cM windows spread throughout the autosomal genome. To circumvent the problem of differential SNP ascertainment biases between population groups, we only considered SNPs with-MAF > 10% in all populations (having corrected for sample size - see Methods). We controlled for heterogeneity in SNP density by first discarding windows containing less than 10 SNPs. The remaining windows containing up to 25 SNPs were thinned to 10 SNPs, and those with 25 SNPs or more were thinned to 25 SNPs. Using the retained SNPs, the number of distinct haplotypes in the 10 SNP windows (H_{10}) and the 25 SNP windows (H_{25}) were estimated separately.

Table 2.1: Estimates of Haplotype Diversity for populations with at least 73 individuals. High values within each continent are shown in bold. Confidence intervals for the haplotype counts are calculated assuming a normal distribution. There were 3,196 distinct 0.5cM windows for the 10 SNP haplotype counts and 2,613 windows for the 25 SNP haplotypes.

Population	H_{10}	**95% Confidence Interval**	H_{25}	**95% Confidence Interval**
Non-Dravidian Influenced	**45.328**	44.711, 45.945	**96.961**	96.241, 97.680
Europe (NW)	40.39	39.835, 40.945	85.555	84.860, 86.251
Europe (NNE)	40.954	40.387, 41.521	86.218	85.523, 86.913
Europe (C)	41.002	40.439, 41.564	86.456	85.755, 87.157
Europe (W)	41.07	40.501, 41.639	87.01	86.308, 87.712
Europe (SE)	41.923	41.345, 42.501	88.702	88.004, 89.401
Europe (SW)	42.64	42.069, 43.212	90.267	89.565, 90.969
Europe (S)	**43.227**	42.637, 43.818	**92.687**	91.964, 93.410
Mexico	**42.345**	41.809, 42.881	**86.967**	86.335, 87.598
Japan	38.274	37.724, 38.824	83.405	82.677, 84.133
Taiwan	**39.698**	39.135, 40.262	**87.382**	86.641, 88.123

Table 2.1 shows the mean and estimated confidence intervals of the distribution of the number of haplotypes. For 10 SNP haplotypes, the East Asian populations have the fewest haplotypes, consistent with a smaller effective population size in East Asia relative to Central Asia and Europe as well as with previous studies of haplotype diversity [Jakobsson et al., 2008, Li et al., 2008, Conrad et al., 2006] and SNP diversity [Keinan et al., 2007]. Interestingly, we find that the Japanese population shows lower diversity than the Taiwanese population. This could be explained either by lower levels of migration or a more severe bottleneck in Japan relative to Taiwan. The Non-Dravidian Influenced group has the highest haplotype diversity of all the sampled populations (and the Dravidian Influenced group shows similar levels of diversity - see Ap-

pendix B), which is the expected pattern if humans migrated out of Africa via the Middle East and into India. The Mexican population has a higher number of haplotypes relative to the East Asian populations, but less diversity than Southern European populations (for both H_{10} and H_{25}). This pattern is consistent (and expected) under a model with East Asian origin of ancestral Native American populations and recent European admixture. Under this model, the initial founder population likely had lower haplotype diversity than the East Asian populations, but European admixture led to increased diversity.

The high number of samples spanning Europe allowed us to investigate geographic patterns of haplotype diversity at a more localized level. We see a north-south gradient in the number of haplotypes present for both H_{10} and H_{25} (Figure 2.3A) with the highest levels of diversity being found in the Southern regions. In particular Southwestern Europe has a higher mean number of haplotypes than Southeastern Europe and Western and Central Europe. This is unexpected, as many current models of historical human migration predict numerous migrations into Europe from Africa via the Middle East, and one would therefore expect the highest diversity in the Southeast, with decreasing diversity moving north and west [Hellenthal et al., 2008, Chikhi et al., 2002, Barbujani and Goldstein, 2004]. The excess haplotype diversity in Southwestern Europe has at least two possible explanations. First, it may reflect direct migration from North Africa across the Mediterranean. Alternatively, it may represent a recolonization of Europe after a period of glaciation during which the Southern areas of Europe became a refugium for the prehistorical human population [Barbujani and Goldstein, 2004, Willis and Whittaker, 2000].

To address this issue, we investigated the level of haplotype sharing between

African and European populations. In the absence of 500K data from North African populations (the HGDP having been genotyped on a different platform), we used the HapMap Yoruba (YRI) population as a proxy for the North African population. Using the 25 SNP haplotype windows outlined above, we found that South West Europe had the highest proportion of haplotypes that are shared withYRI (Table B.4). Furthermore, there were significantly more shared haplotypes between South West Europe and YRI relative to South East Europe and YRI (p-value 3 x 10^{-4}; two-tailed Student's t-test), which suggests that the unusually high haplotype diversity in South Western Europe is indicative of gene-flow across the Mediterranean. However, it is perhaps worth noting that this does not preclude the refugium hypothesis from also contributing to the pattern.

Similarly, we investigated the level of haplotype sharing between Mexico and the European populations. Consistent with historical evidence, the highest proportion of haplotypes in Mexico are shared with South West Europe (Table B.5). However, while the level of haplotype sharing declines from South West Europe, differences between regions do not reach significance. This suggests either incomplete power to detect Mexican haplotypes within Europe, or that European haplotypes are not sufficiently diverged to be isolated to a single region.

Identification of Recent Common Ancestry Runs of homozygosity (i.e. stretches of the genome devoid of heterozygous SNPs) are expected within an individual when both homologous chromosomes share a recent common ancestor. In randomly mating populations, runs of homozygosity may be indicative of historical population demographics, with more runs of homozygosity

expected in populations with a small founder population. Alternatively, long runs of homozygosity (LROHs) are potentially indicative of autozygosity due to recent consanguinity [Li et al., 2006].

To identify LROHs in the POPRES samples, we have developed a method based on a simple hidden Markov model (HMM). The HMM consists of two states for each SNP, which represent either a LROH or a heterozygous region. The emission probabilities at each SNP for each state are dependent on the probability of observing a heterozygote, based on the heterozygosity of the SNP within the population, and the estimated rate of genotyping error. Transition probabilities between the two states are a function of the per-generation recombination rate between SNPs and the (assumed) number of generations since a common ancestor of the two chromosomes. In practice, we call a LROH when the HMM reports the homozygous state as being the most likely state in a region of at least 1cM and containing at least 50 SNPs with a minimum minor allele frequency of 5%. Since hemizygous deletions may also appear as a run of homozygous calls on the genotyping platform, we used GeneChip oligonucleotide hybridization intensities to detect and remove any possible hemizygote deletions from the analysis (see Appendix B).

Examples of the detection of LROH in two individuals are shown in Figure 2.4A. We observe that the majority of individuals exhibit low levels of autozygosity. The median individual in the POPRES sample has approximately 27.6cM of the autosomal genome contained within LROHs (0.8% of the genome, assuming an autosomal map length of 3435.17 cM [Kong et al., 2002]). Furthermore, the median individual within each population shows similarly low levels of autozygosity (Table 2.2). However, the median individuals in Mexico and East

Table 2.2: Long Runs of Homozygosity in individuals, by population.

Population	**cROH in Median Individual**[a] (cM)	**C.I.** [b]	**Individuals with cROH >100cM (%)**	**C.I.** [b]
South Asia	24.12	(23.13, 27.05)	7.5%	(4.2%, 9.6%)
Europe	27.56	(27.40, 28.05)	1.4%	(1.1%, 2.0%)
East Asia	33.87	(31.49, 35.18)	0.0%	n/a
Mexico	47.99	(41.90, 55.72)	5.4%	(1.8%, 9.8%)
All	27.58	(27.23, 27.86)	2.0%	(1.6%, 2.4%)

[a]The cROH in an individual is defined as the total genetic length of all detected long runs of homozygosity at least 1cM in size and containing at least 50 SNPs.
[b]Confidence Intervals calculated by bootstrapping with 1,000 replicates.

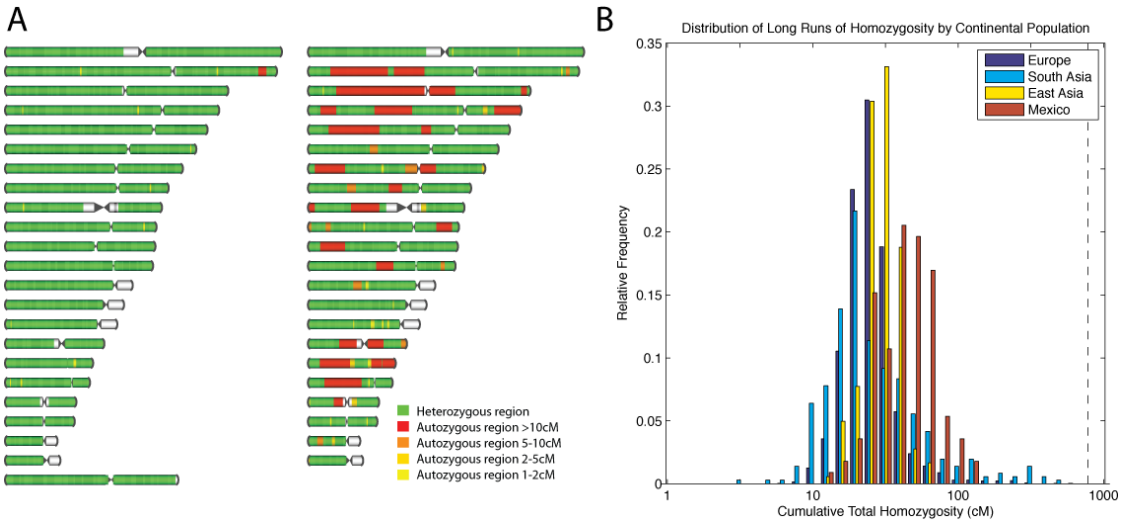

Figure 2.4: Patterns of homozygosity in the human genome. (A) LROHs in two European individuals. The left-hand individual shows typical levels of homozygosity, whereas the right-hand individual shows the mostLROHs in the study. (B) Distribution of individual's cumulative total LROH (shown on a log scale) by continental population. The location of the most extreme individual is indicated by a vertical dashed black line.

Asia have a slightly greater cumulative length of runs of homozygosity (cROH) than the other populations, which most likely reflects smaller founder population sizes in these populations.

While the median individual within in each population has relatively low levels of autozygosity, the distribution of individual levels of autozygosity exhibits long tails, with some individuals showing high levels of homozygosity (Figure 2.4B). All populations have a small number of individuals that are homozygous in over 3% of the genome, and a few individuals are homozygous in over 10% of the genome. These very long runs are suggestive of recent consanguinity. Approximately 2% of individuals in the POPRES survey have more than 100cM of sequence contained in LROH (Table 2.2), and this proportion varies by population (p-value = 3 x 10^{-11}; two-tailed Fisher's exact test).

Certain large regions of the genome appear to be homozygous in a high proportion of individuals. We define Highly Homozygous Regions (HHRs) as regions of at least 50 SNPs which are found to be LROH's in at least 5.0% of individuals within a population. We find 149 HHRs found in at least one population (Figure B.7). While many HHRs are shared across populations, a number are private to a single population with 56, 45, 5, and 3 private HHRs found in Mexico, East Asia, Europe and South Asia respectively. The remaining 40 HHRs are found in more than one population, and 5 are common to all populations. The higher proportion of HHRs in East Asia and Mexico is perhaps indicative of stronger founding bottlenecks for these populations.

Certain HHRs are homozygous in over 10% of POPRES individuals (Table B.8), such as a 2.5Mb region found at 4p15.1, which appears to be a LROH in over 20% of individuals, and the large region around Xq22.3, which is a LROH

in over 32% of East Asians and 12% of Mexicans. We suggest that some of these regions are indicative of recent or ongoing selective sweeps reducing the level of diversity in these region. This would appear to be the case in the region around the lactase gene (*LCT*; 2q21.3), which is known to have undergone a sweep in Europeans [Nielsen et al., 2005]. The region around *LCT* appears as an HHR in Europeans in our study, but not in the other continental populations. Likewise, the *EDAR* gene (2q13), which is related to hair thickness, is known to have undergone a selective sweep in East Asians [Sabeti et al., 2007]. In this study, the region around the *EDAR* gene appears as an HHR in both the East Asian and Mexican populations. However, while a significant number of the remaining HHRs contain genes that have also been associated with sweeps [Williamson et al., 2007], a number of HHRs do not. Why these regions appear homozygotic in significant numbers of individuals remains unknown.

2.4 Discussion

Genome-wide patterns of nucleotide and haplotype diversity within and among human populations can inform our understanding of ancient events in our species history. In contrast, individual genomic patterns are informative of a person's very recent ancestry and can potentially be used to reconstruct personalized genetic history.

In this paper, we have presented a genome-wide study of genotypic and haplotypic variation among 3,845 individuals with ancestry spanning four major geographic regions. As the majority of the data were collected from individuals living in urban areas, it can be considered a cross-section of typical human

genetic diversity in these populations. The data are also of interest due to both the depth of sampling, and the inclusion of two populations (Mexico and South Asia) for which genome-wide genotype data have not been extensively available. The data set is therefore complementary to other large studies of genetic variation such as the HapMap and the HGDP.

Our data provide new insights into the nature of human population structure. The historical admixture between Native American and European populations is clearly visible in the Mexican samples. However, it is perhaps worth noting the relative difficulty in identifying the European region with the highest degree of haplotype sharing with the Mexican individuals. While individuals from South West Europe do share the highest proportion of haplotypes with the Mexican individuals, the difference from other European populations is not statistically significant. As the level of genetic differentiation within Europe is small [Novembre et al., 2008], this is perhaps not surprising. However, with the advent of full sequence data, it will be possible to identify markers that are highly informative of European geographic origin, and hence better understand the history of Mexican population admixture.

To date, few high-density genome-wide SNP studies have been performed in South Asian populations, and the level of genetic diversity in this region is still open to debate [The Indian Genome Variation Consortium, 2008, Rosenberg et al., 2006]. We observe relatively high levels of haplotype diversity in this region, and while regional geographic information was not available for the South Asian individuals in our study, clustering by spoken language suggests the geographic separation is likely an important factor in determining genetic separation in this population as well. Further studies are warranted us-

ing genome-wide data in order to further elucidate the genetic history of this region.

Our analyses also have direct relevance to current debates in human population genetics regarding the extent of historical gene flow among Africa, Europe, and the Middle East [Rando et al., 1998, Simoni et al., 2000, Bosch et al., 2000, Bosch et al., 2002]. Our observation of a north-south gradient in diversity with the highest estimates of diversity in the southern part of the continent is consistent with the initial founding of Europe from the Middle East, the influence of Neolithic farmers within the last 10,000 years, or migrations south followed by a re-colonization of Europe after the Last Glacial Maximum. The unusually high number of haplotypes in South Western Europe is indicative of recurrent gene flow into these regions. Furthermore, when we considered the extent of haplotype sharing with the HapMapYRI population, we found that the South and South-Western subpopulations showed the highest proportion of shared haplotypes. If gene flow had occurred solely through the Middle East, we would expect the South-Easternsubpoupulations to have the highest haplotype diversity and sharing of YRI haplotypes. These two results therefore suggest that while the initial migrations into Europe came via the Middle East, at least some degree of subsequent gene flow has occurred directly over Mediterranean from Africa. Future studies will hopefully be able to better resolve such patterns by comparing haplotypes from further populations around the Mediterranean.

We have also applied a novel method to identify regions of each individual's genome with elevated levels of homozygosity. The vast majority of individuals within the POPRES collection show low levels of homozygosity likely reflective of the recent large effective population size of our species. The small fraction

of individuals who show significantly higher levels of homozygosity are likely offspring of consanguineous unions. Surprisingly, we find a number of regions in the human genome that are homozygous across a high proportion of individuals, and many of these regions are population specific. Though a number of explanations exist for these regions, we anticipate that some are the result of selective sweeps.

The POPRES collection is expected to grow both in terms of the number of individuals within the study, and the number of SNPs genotyped as new genotyping technologies become available. As such we expect that POPRES will become an important resource for ongoing studies in both population and medical genetics.

2.5 Methods

Description of the Data Individuals were genotyped at 500,568 single nucleotide polymorphisms (SNPs) on the Affymetrix GeneChip Mapping Array 500K set byGlaxoSmithKline as part of the POPRES initiative [Nelson et al., 2008]. As such, a full description of the sampling protocol can be found in [Nelson et al., 2008]. Briefly, individuals were sampled in 8 batches between November 2005 and March 2007. In total, a total 6 studies contributed to the POPRES samples studied in this paper, details of which can be found in [Nelson et al., 2008]; TheCoLaus study (2,508 individuals sampled in Lausanne, Switzerland), the LOLIPOP study (843 individuals sampled in London, England), Healthy Caucasian Controls (201 individuals sampled in Adelaide, Australia, North Carolina, USA or Ottawa, Canada), Healthy Mexican Controls

(112 individuals sampled in Guadalajara, Mexico), Healthy Taiwanese Controls (108 individuals sampled in Taipei, Taiwan) and Healthy Japanese Controls (73 individuals sampled in Sydney, Australia).

Individuals in the CoLaus study (covering individuals of European and South Asian ancestry) were asked for information regarding parental and grand-parental country of birth. Primary language information was also collected for sections of theCoLaus and LOLIPOP studies. For certain analyses detailed in this paper, we used a "strict" dataset of 2,943 individuals which excludes individuals with either ambiguous or reported mixed ancestry, PCA outliers, and those with estimated identity by descent with another individual in the sample greater than 20%.

European individuals were assigned to countries using grand-parental country of birth where possible. If all observed grandparents originate from a single country, then that country was used as the ancestral location for the individual. In the case of mixed ancestry, individuals were assigned to a separate group (Mix). In the absence of grand-parental information, individuals were assigned on the basis of country of birth. Mexican and East Asian individuals were assigned to groups on the basis of self-identified ancestry. Finally, South Asian individuals were assigned to groups on the basis of spoken language. Full details of the ancestral data available, and assigned groupings, for each individual is available as a supplementary table. For certain subsequent analyses, country and language groupings were combined to form larger groups, as detailed in Table B.2.

SNP positions were mapped to NCBI build 36.1 (UCSC hg18). After applying quality control filters [Nelson et al., 2008], a total of 443,434 SNPs re-

mained, giving an average SNP spacing of 1 SNP every 6.4kb in the assembled genome. Individuals have an average missing genotype rate of approximately 2.3%. Summaries of minor allele frequency spectra are given in the supplementary material.

Principal Component Analysis Principal Component Analysis (PCA) was conducted using the program *smartpca* contained in version 2.0 of the *Eigensoft* package [Patterson et al., 2006]. The analysis was run without the removal of outliers. To avoid artifacts due to linkage disequilibrium we first used *PLINK* [Purcell et al., 2007] to thin the data by excluding SNPs with pairwise genotype $r^2 > 0.8$ within a sliding window of 50 SNPs.

For the global PCA analysis, we combined the POPRES dataset with 479 individuals from the HGDP [Jakobsson et al., 2008]. As the two datasets were obtained using separate genotyping platforms, only a subset of SNPs are common to both. After requiring that no SNP have more than 5% missing data, and removing SNPs in highLD as described above, the combined dataset consisted of 73,520 SNPs in 3,448 individuals.

For the Asian sub-continental PCA analysis we excluded related individuals and kept 271 individuals from Japan, Taiwan, and HapMap JPT and CHB. In our analysis of regional structure within South Asia, we included 315 individuals from India and Sri Lanka whose language information was known and not primarily English. For the Mexican admixture analysis, we created a set of 778 individuals which includes the Mexican individuals, a small subset of mainland Europeans used in the world *STRUCTURE* analysis, and East Asian populations used in the Asian sub-continental analysis.

***STRUCTURE* Analysis** For the global *STRUCTURE* analysis, we attempted to reduce the effect of sample size by using a reduced dataset of 1,245 individuals from the strict dataset (and including the HapMap samples). Due to the large number of European samples, we include only individuals of with self-reported ancestry from mainland European countries. Any country with more than 15 individuals was reduced to 15 individuals, selected at random from the population. We also exclude individuals found to be outliers based on preliminaryPCA runs conducted separately on the East Asian, European, and South Asian samples. PCA-based outliers were determined by using *smartpca* with default settings [Patterson et al., 2006]. This approach removes individuals whose PC coordinates are more than 6 standard deviations from the mean coordinate along any of the top 10 principal components, and repeats this process for a maximum of 5 iterations.

In order to make the run-time tractable, we reduced the number of markers to 6,567 SNPs selected to have MAF > 0.2 and a minimal separation of 400kb. We ran *STRUCTURE* version 2.2 without prior population assignment, using the correlated alleles model, with 10,000 iterations burn-in and 10,000 run time. We used the INFERALPHA option under the admixture model (also known as the F model), with the allele frequency prior parameter LAMBDA set to 1. Results were plotted using *Distruct* [Rosenberg, 2004]. The results for $K = 2$ to $K = 6$ are shown in Figure B.2A. Repeated runs of *STRUCTURE* give qualitatively similar results.

The sub-continental analyses are described in Appendix B using a similar SNP selection method with markers selected independently for each analysis.

Haplotype Diversity As described in the main text, we summarized haplotype diversity by the number of distinct haplotypes contained within 0.5 cM windows. Haplotypes were obtained using *BEAGLE* [Browning and Browning, 2007], as described in the Appendix B. While the amount of ascertainment bias for the Affymetrix 500K chips is difficult to characterize, it is likely to vary from population to population. In order to circumvent the problem of ascertainment bias, we only considered SNPs withMAF > 10% in all of the studied population groups (after sample size correction - see below). By only including the SNPs common to all populations in the diversity analyses, differences in haplotype diversity among populations are largely governed by differences in the effective population size between populations.

Using the Phase II HapMap genetic map [Frazer et al., 2007], we divided the genome into 0.5 cM windows. For each chromosome, the first window started at the position of the first SNP and then extended 0.5 cM downstream. The second window started at the position where the first window ended, regardless of SNP locations. To ensure that separate regions of the genome had similar numbers of SNPs for the estimation of haplotype diversity, we selected a subset of SNPs within each window. We classified each of the 0.5cM windows into one of three groups: 1) < 10 SNPs, 2) 10-24 SNPs, 3) ≥ 25 SNPs. Windows having < 10 SNPs were excluded from the analysis, as it was likely that haplotype diversity would be low in all populations. For windows having 10-24 SNPs, we selected a random sub-set of 10 SNPs for each window. For windows with ≥ 25 SNPs, we selected a random sub-set of 25 SNPs for each window. For each window, the same set of SNPs was chosen for all of the population groupings. In the subsequent analyses, the windows with 10 SNPs were analyzed separately from the windows with 25 SNPs.

The number of haplotypes in each region of the genome is confounded by the number of chromosomes sampled in each population, as populations with more sampled chromosomes will be more likely to include rare haplotypes. As the number of individuals sampled from each population varies quite dramatically (Table B.2), we selected a random sub-set of 73 individuals from each population to use for subsequent analyses. Populations with samples sizes below 73, namely the Dravidian Influenced and Europe ESE groups, were excluded from the analysis. Minor allele frequencies were calculated for each SNP in each population using these smaller sub-samples of 73 individuals from each population. We repeated the haplotype analyses using several different random sub-sets of 73 individuals and did not see a substantial difference between replicates.

Identification of Runs of Homozygosity To identify runs of homozygosity, we developed a novel method based on a Hidden Markov Model (HMM). The model consists of two hidden states, namely autozygous (A) and non-autozygous ($\neg A$). If SNP i has genotypic state X_i (= 0, 1, 2, where 1 is the heterozygous state), and hidden state S_i, the emission probabilities for the two states at each SNP are given by:

$$\begin{aligned}
\Pr(X_i = 1 | S_i = \neg A) &= h \\
\Pr(X_i = (0,2) | S_i = \neg A) &= 1 - h \\
\Pr(X_i = 1 | S_i = A) &= \varepsilon \\
\Pr(X_i = (0,2) | S_i = A) &= 1 - \varepsilon
\end{aligned}$$

where h is the observed SNP heterozygosity in the population, and ε is the assumed genotyping error. We set $\varepsilon = 0.2\%$.

The transition probabilities between hidden states are a function of the genetic map distance between SNPs, as estimated by the Phase II HapMap [Frazer et al., 2007], and the expected number ofmeioses (M) since a recent common ancestor.

$$\begin{aligned}
\Pr(S_{i+1} = A | S_i = \neg A) &= \Pr(S_{i+1} = A)\left(1 - e^{-2M(r_{i+1} - r_i)}\right) \\
\Pr(S_{i+1} = A | S_i = A) &= 1 - \Pr(S_{i+1} = A | S_i = \neg A) \\
\Pr(S_{i+1} = \neg A | S_i = A) &= \Pr(S_{i+1} = \neg A)\left(1 - e^{-2M(r_{i+1} - r_i)}\right) \\
\Pr(S_{i+1} = \neg A | S_i = \neg A) &= 1 - \Pr(S_{i+1} = \neg A | S_i = A)
\end{aligned}$$

where r_i is the genetic map location of SNP i in Morgans. In practice, we chose M to be 4 to reflect our interest in homozygosity caused by recent common ancestry. However, we have found the method to be largely robust to values of M up to 10 (data not shown). For the prior probabilities of being in theautozygous or non-autozygous state, we chose 0.05 and 0.95 respectively.

We use the Viterbi algorithm to find the most likely hidden state path (see, for example, [Durbin et al., 1999]).

Acknowledgements

We dedicate this paper to the memory of our friend and colleague, Scott Williamson. We thank Brian Browning for providing the program *BEAGLE* and advice for phasing. We also thank Jonathan Pritchard for making the *STRUCTURE* source code available. Andy Clark, Hong Gao, Ryan Hernandez, Sean Myles, and Keyan Zhao provided numerous helpful insights. This work was supported by NIH R01GM083606 (CDB,KB), NIH 1U01HL084706 (ARB, AI),

an NSF graduate research fellowship (KEL), NSF-DBI 0701382 (AR,MHW), and NSF 0516310 (JD).

Author Contributions

Performed analyses: AA, KB, ARB, KL, JD, RNG. Conceived analyses: AA, KB, ARB, KL, JN, CDB. Contributed data: KSK, MRN. Quality control of data: ARB, MHW, AI, AR. Wrote the paper: AA, CDB.

CHAPTER 3

GENOME-WIDE PATTERNS OF POPULATION STRUCTURE AND ADMIXTURE IN WEST AFRICANS AND AFRICAN AMERICANS*

*Originally published as: K. Bryc, A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser, S. Williams, A. Froment, J.-M. Bodo, C. Wambebe, S. A. Tishkoff, and C. D. Bustamante (2010). *Proc Natl Acad Sci*,107(2):786 - 791

3.1 Abstract

Quantifying patterns of population structure in Africans and African Americans illuminates the history of human populations and is critical for undertaking medical genomic studies at a global scale. To obtain a fine-scale genome-wide perspective of ancestry, we analyze Affymetrix 500K genotype data from African Americans (n = 365) and individuals with ancestry from West Africa (n = 203 from 12 populations) and Europe (n = 400 from 42 countries). We find that population structure within the West African sample reflects primarily language and secondarily geographic distance, echoing the Bantu expansion. Among African Americans, analysis of genomic admixture by a principal component based approach indicated that the median proportion of European ancestry is 18.5% (Inter Quartile Range: 11.6%, 27.7%) with very large variation among individuals. In the African American sample as a whole, few autosomal regions showed exceptionally high or low mean African ancestry, but the X-chromosome was predominantly of African origin, consistent with a sex-biased pattern of gene flow with an excess of European male and African female ancestry. We also find that genomic profiles of individual African Americans afford personalized ancestry reconstructions differentiating ancient vs. recent European and African ancestry. Finally, patterns of genetic similarity among inferred African segments of African American genomes and genomes of contemporary African populations included in this study suggest African ancestry is most similar to non-Bantu Niger-Kordofanian speaking populations, consistent with historical documents of the African diaspora and trans-Atlantic slave trade.

3.2 Introduction

Studies of African genetic diversity have greatly informed our understanding of human origins and history [Reed and Tishkoff, 2006, Tishkoff et al., 2009], identified genes under natural selection across evolutionary time [Tishkoff et al., 2007] and hold great potential for elucidating the genetic bases of disease susceptibility and drug response among diverse human populations [Sirugo et al., 2008, Campbell and Tishkoff, 2008]. The study of African population structure is also critical for reconstructing patterns of African ancestry among African Americans and for enabling genome-wide association mapping of complex disease susceptibility and pharmacogenomic response in African American populations [Ma et al., 2005, Williamson et al., 2000, Reich et al., 2005, Johnson, 2008].

Africa contains over 2000 ethno-linguistic groups and harbors great genetic diversity [Tishkoff et al., 2009, Frazer et al., 2007, Garrigan et al., 2007, Jakobsson et al., 2008, Li et al., 2008, Tishkoff et al., 1996], but little is known about fine-scale population structure at a genome-wide level. This is, in part, because previous studies of high-density SNP and haplotype variation among global human population (defined as studies with at least 100,000 Single Nucleotide Polymorphism markers) have included few African populations [Frazer et al., 2007, Jakobsson et al., 2008, Li et al., 2008, Adeyemo et al., 2005] while detailed studies of genetic structure among African populations have utilized a modest number of markers [Tishkoff et al., 2009, Tishkoff and Kidd, 2004, Tishkoff and Verrelli, 2003, Tishkoff and Williams, 2002] (approximately 1,500 microsattelites and indels). Nonetheless, recent studies of microsatellite and DNA sequence variation suggest significant population structure exists within sub-Saharan Africa

with geography, language, and mode of subsistence (e.g., hunter-gatherer, pastoralist, agriculturalist) as potential key factors [Tishkoff et al., 2009, Jakobsson et al., 2008, Li et al., 2008, Patin et al., 2009]. Given that high-density genotype data have revealed discernible population structure within other continental populations (e.g., Europe and East Asia) and even among geographic regions within countries (e.g., Switzerland, Finland, United Kingdom) (e.g., [Lao et al., 2008, Novembre et al., 2008, McEvoy et al., 2009, Nelis et al., 2009]), there is strong reason to believe that high-density genotype data from African and African-American populations can further elucidate patterns of genetic structure among these important populations.

We have, therefore, genotyped on the Affymetrix 500K gene chip more than 200 individuals from 11 populations in West and South Africa (Table C.1; Figure C.1) who speak Nilo-Saharan, Afro-Asiatic and Niger-Kordofanian languages and integrated these data with our previous studies of human genomic diversity including 365 African-Americans from throughout the United States and 400 individuals of European ancestry [Frazer et al., 2007, Nelson et al., 2008]. We used Principal Component Analysis (PCA) to infer axes of genetic variation within Africa, and examined individual and population clustering using the clustering algorithm FRAPPE [Tang et al., 2005]. For each African American subject, we have also evaluated individual patterns of European and African ancestry along each chromosome using a novel and computationally efficient PCA-based method that infers admixture proportions based on high-density, genome-wide data.

3.3 Results

Genetic Structure of West African Populations. Our study focused on West African populations, since previous genetic and historical studies suggest that region was the source for ancestry of present-day African-Americans. Among these West African populations, Wrights measure of population differentiation (autosomal F_{ST} [Weir and Cockerham, 1984]) was low (1.2%), suggesting quite recent common ancestry of all individuals in our sample. Nonetheless, we observed substantial variation in pairwise F_{ST} among populations, suggesting genetic heterogeneity among the groups (see Table 3.1).For example, The Fulani appear to be genetically distinct from all other West African populations we sampled (average pairwise F_{ST} = 3.91%). Likewise, we found that the Bulala, Xhosa, and Mada populations consistently exhibited pairwise F_{ST}'s above 1% when compared to any other population, while the non-Bantu Niger-Kordofanian populations of the Igbo, Brong, and Yoruba exhibited little genetic differentiation from one another (average $F_{ST} < 0.4\%$). These results suggest that there are clear and discernible genetic differences among some of the West African populations while others appear to be nearly indistinguishable even when comparing over 300,000 genetic markers.

In order to investigate whether we could reliably distinguish ancestry among individuals from these populations, we utilized two approaches tailored for high-density genotype data. The first, *FRAPPE*, implements a maximum likelihood method to infer genetic ancestry of each individual, where the individuals are assumed to have originated from K ancestral clusters [Tang et al., 2005]. Figure 3.1A and Figure C.2 summarizes *FRAPPE* results when the number of clusters, K, is varied from $K = 2$ to $K = 7$. The small

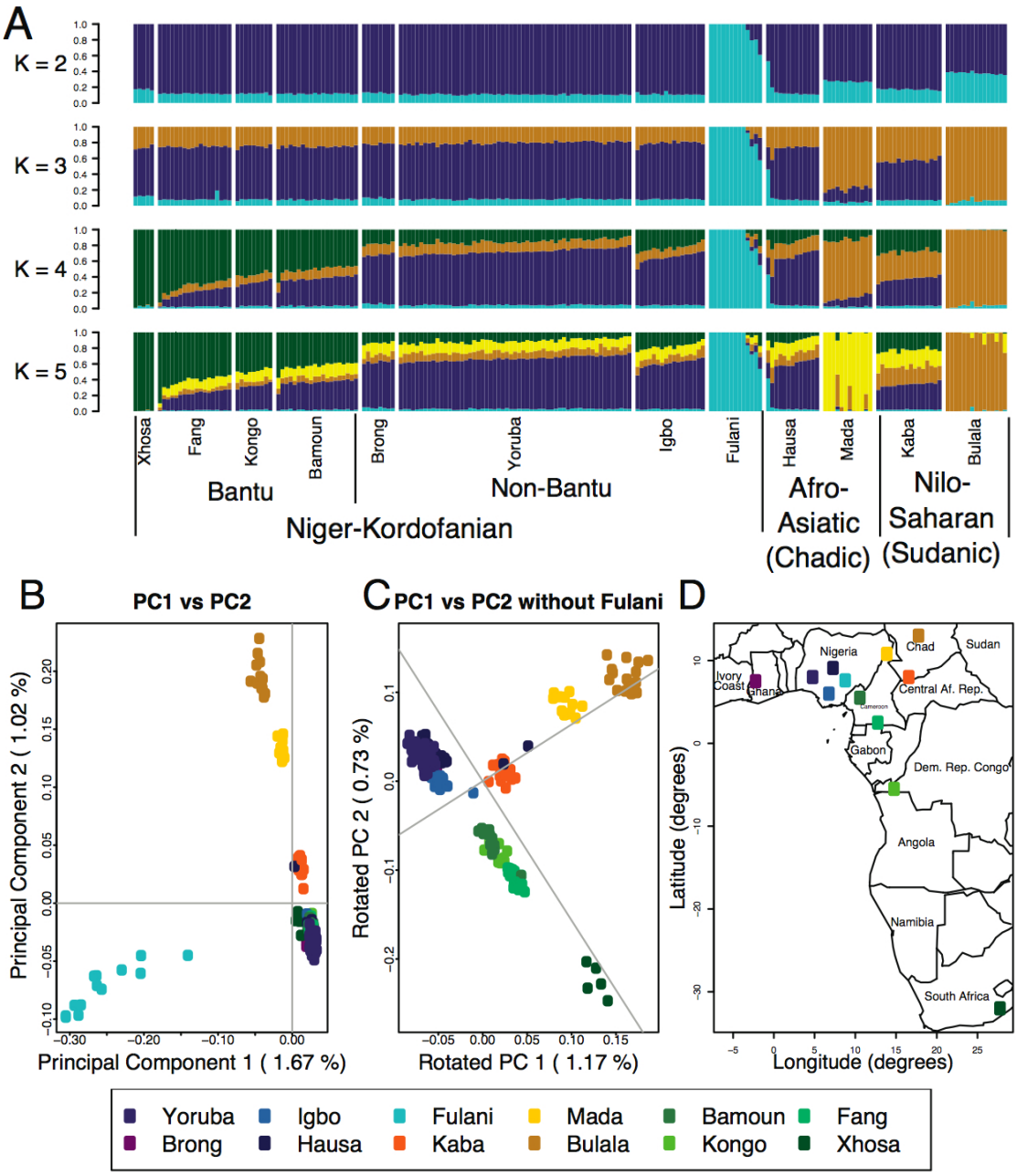

Figure 3.1: Population structure within West Africa and relation to language and geography. A) *FRAPPE* analysis of the African populations. Individuals are represented as thin vertical lines partitioned into segments corresponding to the inferred membership in $K = 2$ through $K = 5$ genetic clusters as indicated by the colors. B) Principal components 1 and 2 of the African individuals. C) Principal components 1 and 2 of the African individuals, excluding the Fulani population where the components have been rotated to further emphasize similarity with geography. D) Approximate locations of sampled populations in Africa.

Table 3.1: F_{ST} distances among African populations.

	Igbo	Brong	Yoruba	Kongo	Bamoun	Xhosa	Fang	Hausa	Kaba	Mada	Bulala
Brong	0.350%	-									
Yoruba	0.084%	0.200%	-								
Kongo	0.282%	0.425%	0.291%	-							
Bamoun	0.293%	0.448%	0.318%	0.175%	-						
Xhosa	1.448%	1.636%	1.251%	1.106%	1.277%	-					
Fang	0.415%	0.594%	0.432%	0.150%	0.247%	1.165%	-				
Hausa	0.397%	0.560%	0.420%	0.588%	0.546%	1.796%	0.691%	-			
Kaba	0.516%	0.510%	0.471%	0.501%	0.484%	1.498%	0.567%	0.619%	-		
Mada	1.296%	1.336%	1.300%	1.282%	1.276%	2.319%	1.380%	1.299%	0.968%	-	
Bulala	1.862%	1.905%	1.879%	1.736%	1.806%	2.646%	1.929%	1.773%	1.280%	0.931%	-
Fulani	3.905%	3.684%	4.034%	3.770%	3.996%	4.133%	4.063%	3.761%	3.811%	3.967%	3.920%

[1]The cROH in an individual is defined as the total genetic length of all detected long runs of homozygosity at least 1cM in size and containing at least 50 SNPs.
[2]Confidence Intervals calculated by bootstrapping with 1,000 replicates.

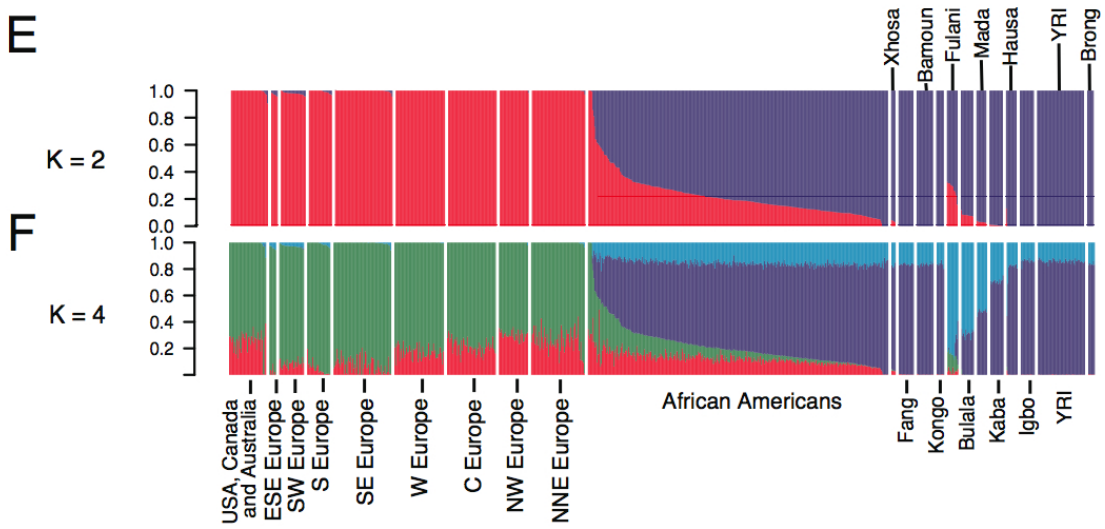

Figure 3.2: *FRAPPE* analysis of Europeans, Africans, and African Americans. (E, F) *FRAPPE* clustering of Europeans, African Americans, and Africans. Individuals are represented as thin vertical lines partitioned into K segments corresponding to the inferred membership of the genetic clusters indicated by the colors. Values for $K = 2$ (E) And $K = 4$ (F) are shown for comparison between the two analyses.

number of clusters was consistent with the small, overall, level of population differentiation among these populations. We next undertook principal component analysis of the matrix of individual genotype values (i.e., the matrix with entries 0, 1, or 2 generated by tallying the number of copies of a given allele across all SNPs in a panel for all individuals genotyped) [Patterson et al., 2006].

Patterns of population structure were consistent between the two approaches (Figure 3.1). For example, in the *FRAPPE* analysis, the Fulani population was distinguished at $K = 2$, with Bulala, Mada and Kaba showing some shared ancestry with the Fulani. Principal Component Analysis, likewise, separated the Fulani from other populations along the first principal component (PC1) (Fig. 3.1B). The two subsequent principal components, PC2 and PC3, reflect the geographical distribution of the populations. PC2 showed a Chadic and Nilo-Saharan dimension extending into inland Africa from the coast, distinguishing the Bulala, Mada, and Kaba populations. These populations belong to the Nilo-Saharan and Afro-Asiatic (Chadic) linguistic groups, and live further inland. Analysis of the African populations excluding the Fulani gave PC1 and PC2 that resemble the second and third principal components of the PCA with the Fulani (Fig. 3.1C). Rotating the PC1 and PC2 axes from the PCA without the Fulani reveals the similarity of the genetic and geographic maps (Fig. 3.1C and 3.1D).

At $K = 3$, the *FRAPPE* algorithm clusters the Bulala into their own group and suggests shared ancestry among the Mada, Kaba, and Hausa potentially indicating differentiation of Nilo-Saharan and Afro-Asiatic speaking populations from Niger-Khordofanian speaking populations. At $K = 4$, all individuals from the Bantu-speaking Xhosa of South Africa cluster into a single group and

individuals from the Bantu-speaking populations (Fang, Bamoun and Kongo) exhibit considerable shared membership in this cluster. At $K = 5$ the Mada are distinguishable as a unique group, with modest shared ancestry with the Hausa and Kaba as well as most of the Niger-Kordofanian populations. These results suggest that while these populations are quite closely related genetically, it is possible to detect meaningful population substructure given sufficient marker density (see also [Tishkoff et al., 2009]). In order to compare patterns of haplotype structure and discern differences in demographic history among the African populations, we estimated linkage disequilibrium (LD) between all pairs of markers in the data for all populations (see Figure C.3). All of the African populations showed low levels of linkage-disequilibrium (even at closely linked sites) and a rapid decay of LD with distance genome-wide relative to populations of European ancestry, which could potentially affect the ability to accurately infer haplotypes in these populations.

Genome wide patterns of admixture in African Americans. To better understand the genetic structure of the African American population and to determine African American ancestry, we used *FRAPPE* to evaluate African Americans together with European and African individuals genotyped on the same marker set. At $K = 2$, African populations (blue) were distinguished from European populations (red), with African Americans showing highly variable levels of European and African ancestry (Figures 3.2E, 3.2F). For the African Americans, estimated mean African ancestry was 77%, consistent with prior studies [Tishkoff et al., 2009, Parra et al., 2001, Salas et al., 2005, Lind et al., 2007, Smith et al., 2004, Parra et al., 1998]. Analysis at $K = 4$ revealed additional substructure in a North-South cline within Europe and clusters coin-

ciding with the linguistic and geographic substructure within Africa (see Appendix C and Figures C.4 and C.5 for additional *FRAPPE* and population genetic analyses). Principal Component Analysis of the genotype value matrix of the European, West African, and African-American samples revealed the primary axis of variation (PC1) to correspond with European vs. African ancestry (see Figure 3.3A) and explain approximately 9.8% of the genetic variance. Specifically, we observed two centroids in the data with the all individuals of European ancestry exhibiting negative loadings along PC1 while all the West African individuals exhibited positive loadings. African-Americans exhibited a wide range of loadings along PC1 presumably due to differences in European versus African ancestry. The second principal component (PC2) corresponds to population substructure within West Africa and largely mirrors the patterns discussed above.

Estimation of admixture in local genomic regions. We reconstructed estimated European or African ancestry for every African-American in our data set at every position in the genome using a PCA based algorithm (Figure 3.3A). Our method is a generalization of the Paschou et al. approach [Paschou et al., 2007] and estimates genome-wide proportion of African ancestry for a given individual as p = b/(a + b) where b and a are the chord distances from the European and African centroids, respectively, for the given individual along the first principal component. Our generalization involves undertaking the PC1 distance analysis on a grid of points along the genome (as opposed to genome wide) centered on 15 SNP windows and using a Hidden Markov Model for inference of ancestry state (i.e., having 0, 1, or 2 chromosomes of recent African origin; see Figure 3.3B, Materials and Methods, Appendix C, and Figure C.6). An ancestry

plot summarizing the number of segments of European (i.e., 0), African (i.e., 2), or admixed (i.e, 1) ancestry for a representative African American individual with 73.5% African ancestry is illustrated in Figure 3.4C. There is a great deal of variation among the ancestry plots of the self-identified 365 African Americans in the study, ranging from an estimate of over 99% African ancestry to an estimate of less than 1% African ancestry (Figure 3.4F). Some patterns reflected a high level of African ancestry and only one or two ancestry-informative events per chromosome, suggesting very recent direct African ancestry (Figure 3.4D). Other patterns reflected only European and admixed ancestry throughout the genome, suggesting one parent of European ancestry and one parent of African American ancestry (Figure 3.4E).

An interesting question one can address with these kinds of data is whether regions of the genome show substantially high European or African ancestry across all individuals in the sample (as may be the case, for example, if a particular allele from one of the ancestral populations was under strong selection [Workman et al., 1963, Reed, 1969, Cavalli-Sforza and Bodmer, 1971, Tang et al., 2007]). For our analysis, we considered genomic regions as potential candidates for increased European or African ancestry if the mean ancestry for the region across the 365 African Americans individuals was 3 standard deviations above or below the genome wide average of African ancestry (78.1%). Using this approach, we found that several genomic regions of autosomal chromosomes 5, 6, and 11 could be considered outliers from the genome-wide distribution of ancestry (Figure 3.5), although these differences were not significant after correction for multiple tests (Figure C.7, Table C.2). In contrast to the autosomes, the X chromosome shows significantly high African ancestry along the majority of the chromosome, consistent with a sex-biased model of admixture

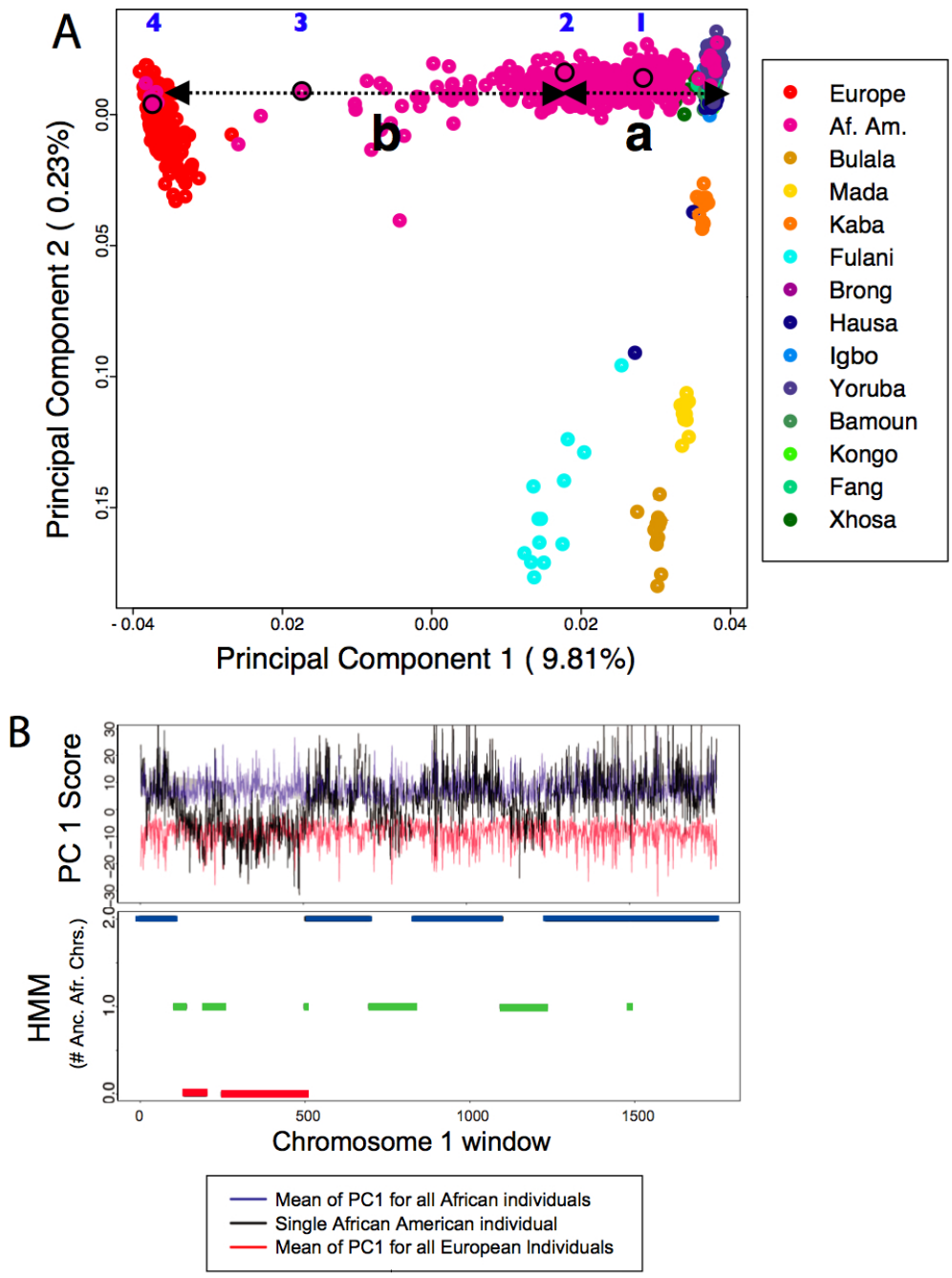

Figure 3.3: Illustration of our PCA based ancestry estimation method. A) Graphical illustration of approach: Euclidean distances from a given individuals coordinates in PCA space (i.e., loadings) and the African centroid (a) and the European centroid (b) along PC1 for PCA space that includes only Europeans, African Americans, and West Africans. B) Local ancestry estimation using PCA sliding window approach and associated Hidden Markov Model (HMM) for number of chromosomes for a given individual (i.e., 0,1, or 2) with African ancestry.

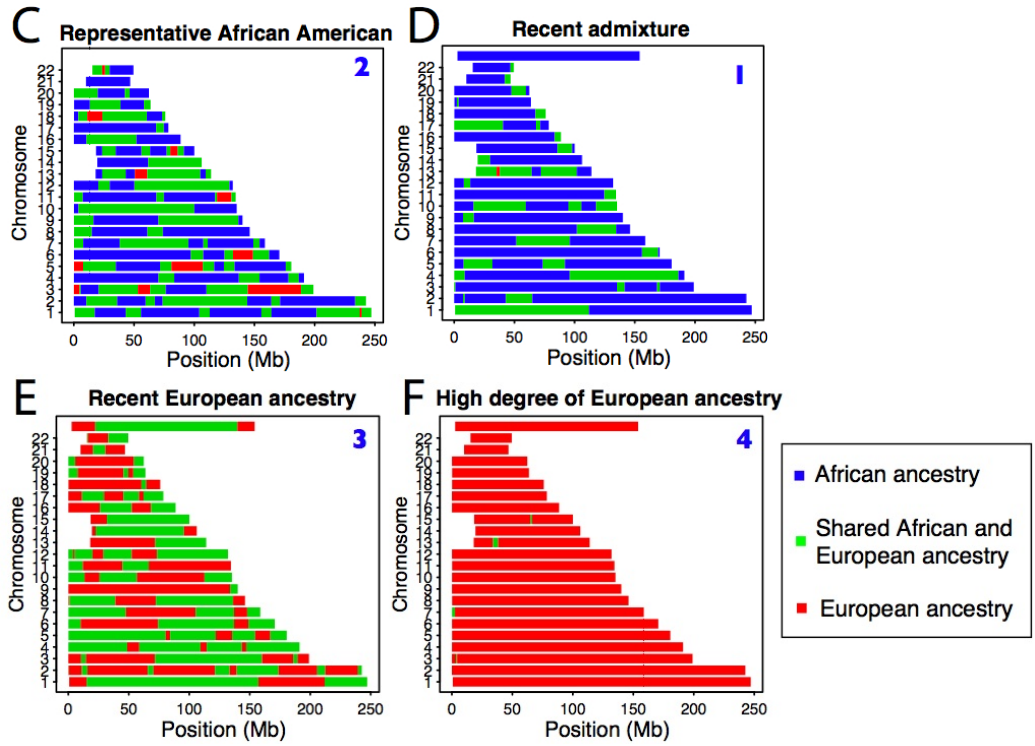

Figure 3.4: Individual ancestry results of our PCA based ancestry estimation method. (C, D, E, F) Individual ancestry estimates of four representative African American individuals in our data set of 365. The colors represent two chromosomes of African ancestry (blue), two chromosomes of European ancestry (red), or one chromosome of African and one chromosome of European ancestry (green). G) Mean ancestry of 365 African American individuals at each window across chr 1, chr 11, chr 12, and chr X. The black line shows the overall mean estimated ancestry. Red bands indicate +3 and -3 standard deviations from the mean ancestry.

with excess European male and African female ancestry (Figure 3.5).

3.4 Discussion

The Bantu expansion occurred approximately 4,000 years ago, originating in Cameroon or Nigeria and expanding throughout Sub-Saharan Africa. The clustering of the Xhosa, Fang, Bamoun and Kongo populations, all of which

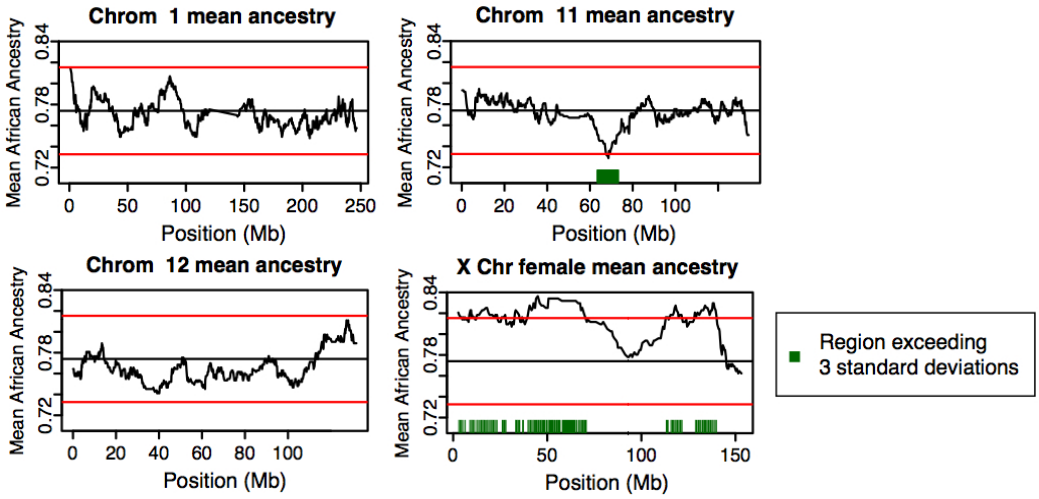

Figure 3.5: Mean ancestry ancestry of 365 African American individuals at each window across chr 1, chr 11, chr 12, and chr X. The black line shows the overall mean estimated ancestry. Red bands indicate +3 and -3 standard deviations from the mean ancestry.

are Bantu-speaking Niger-Kordofanian populations, likely reflects a Bantu migration from Nigeria/Cameroon expanding towards the South [Ehret, 2001, Klieman, 2003]. The relative order of clustering (first the East-West axis followed by the North-South axis) suggests that the strongest differentiating axis among these populations is linguistic classification corresponding to Chadic and Nilo-Saharan versus Niger-Kordofanian ancestry. The relatively weaker North-South axis may result from the genetic similarity among the Niger-Kordofanian linguistic groups due to their recent common ancestry from a proposed homeland in Nigeria/Cameroon. Although sampled in Nigeria, the very distinct Fulani are a part of a nomadic pastoralist population which occupies a broad geographic range across central and western Africa. Analyses of microsatellite and insertion/deletion polymorphisms indicate that they share ancestry with Niger-Kordofanian, North African, and central African Nilo-Saharan populations, as well as low levels of European and/or Middle Eastern ancestry [Tishkoff et al., 2009]. Exempting the Fulani, our LD analyses show no

large differences in rates of LD decay among our sampled African populations with all populations have a faster decay of LD (i.e., larger inferred effective population size) than previously characterized populations of European ancestry (see Appendix C).

Interestingly, the Kongo population does not follow the overall trend of East-West and North-South clustering. The Kongo populations genetic proximity to geographically distant Bantu populations could be explained by the genetic similarity of Bantoid speaking populations in the region, as seen in the FRAPPE analyses (Figure 3.1). Alternatively, while these individuals self-identified as Kongo and were refugees from locations within the Democratic Republic of Congo, the samples were collected in Cameroon, and therefore self-identified ancestry might poorly represent the long-term geographic origins, or may reflect recent admixture.

A concern in estimating admixture is the effect of choice of ancestral populations – often, the true ancestral population is no longer available for sampling, so using a proxy may introduce bias when evaluating the admixed population. For example, individual admixture estimates in Latin Americans have been shown to depend on the ancestral populations evaluated [Tian et al., 2008]. Most studies estimating admixture proportions in African Americans have used a single ancestral African population, the Yoruba, [Tang et al., 2007], and our data provide an effective means of testing whether other populations may serve as better proxies for the ancestral population of African-Americans and whether using the Yoruba biases inferences. Comparison of the inferred African segments of African-American genomes to contemporary African populations (see Table C.3) reveals that the ancestry of the African component of African Amer-

icans is similar to the non-Bantu Niger-Kordofanian profile which includes the Igbo, Brong, and Yoruba, with F_{ST}'s to African segments of the African Americans ranging from 0.074% to 0.089%. That these F_{ST} values are all nearly identical (and quite small) coupled with the small pairwise F_{ST} of the Igbo, Yoruba, and Brong populations (Table 3.1), suggests that any of these populations may serve as a good proxy for the ancestral population of the African Americans and that, in fact, all three likely contributed ancestry to present day African-Americans. This is wholly in line with historical documents showing that the Igbo and Yoruba are two of the ten most frequent ethnicities in slave trade records [Hall, 2005].

That some individuals who self-identify as African American show almost no African ancestry, while others show almost complete African ancestry has implications for pharmacogenomics studies and assessment of disease risk. Although individuals with very low African or very low European ancestry may be expected by chance after several generations of admixture, these individuals are most likely descendants of individuals of European ancestry or recent African immigrants, respectively. Assuming these individuals are not simply mislabeled, it appears that the range of genetic ancestry captured under the term African American is extremely diverse, which suggests caution should be used in prescribing treatment based on differential guidelines for African Americans [Reiner et al., 2005].

We found regions on chromosomes 5, 6, and 11 that show deviations from the overall mean African ancestry. These regions do not overlap with those previously suggested to be under selection [Tang et al., 2007] and about a dozen

genes are found across these regions. Whether these genes or regions are potentially under selection in African Americans merits further investigation.

In conclusion, we believe the data presented here speak to several important points. First, patterns of genomic diversity within Africa are complex and reflect deep historical, cultural, and linguistic impacts on gene flow among populations. These patterns are discernible using high-density genotype data and allow us to differentiate closely related populations along linguistic and geographic axes. Secondly, admixture can be reconstructed for local genomic regions efficiently at a high density of genetic markers. For this study, we tailored the method to admixed populations with two ancestral source populations, but the approach is generalizable. Application of the method to genome-wide patterns of genomic variation in African Americans reveals the rich mosaic structure of admixture in this population. We find that we can distinguish African ancestry among West African populations to a large degree (e.g., Bantu from non-Bantu Niger-Kordofanian profiles), but that some populations (e.g., the Igbo, Yoruba, and, to a lesser extent, Brong) are so closely related genetically that their contribution to patterns of African ancestry in African-Americas is not reliably distinguishable. We believe that increasing the density of markers and, more importantly, sequencing directly in these populations to identify ancestry informative markers may in the future make this possible.

3.5 Methods

Datasets. We genotyped 225 individuals from 11 African populations (see [Tishkoff et al., 2009] for sampling locations) on the Affymetrix 500K array set,

and incorporated data from the Yoruban population of Ibadan, Nigeria from the HapMap project, thinned to the same SNP set [Frazer et al., 2007]. European samples were from the GlaxoSmithKline POPRES project, a resource of nearly 6,000 control individuals from North America, Europe, and Asia [Nelson et al., 2008] genotyped on the Affymetrix GeneChip 500K array set. We extracted for our analyses a subset of 400 individuals from Europe, randomly sampling 15 individuals each European country represented in POPRES where possible, and 15 individuals each from the USA, Canada, and Australia. We include 365 African Americans from this dataset (see also Appendix C and [Nelson et al., 2008]). The use of these data is consistent with written informed consent provided by the study participants and approved by the proper Institutional Review Boards, and permits were obtained for collection of African populations as described in [Tishkoff et al., 2009].

Population structure analyses. *FRAPPE* implements an efficient maximum likelihood version of Bayesian clustering algorithm, *STRUCTURE* [Tang et al., 2005, Pritchard et al., 2000, Falush et al., 2003]. After thinning markers to have Pearson product-moment correlation of allele frequency, r^2, less than 0.5 in 50 SNP windows, shifted and recalculated every 5 SNPs, we ran *FRAPPE* on all 204,457 remaining markers for 5,000 iterations. Clusters at $K = 6$ and higher did not correspond to known linguistic or population substructure (see Figure C.2). We ran PCA using the program smartpca from the package *eigenstrat* [Patterson et al., 2006] on a reduced dataset of 251,253 SNPs where $r^2 < 0.8$ in 50 SNP windows. F_{ST} was calculated using a C++ implementation of Weir and Cockerhams F_{ST} weighted equations from [Weir and Cockerham, 1984]. Minor allele frequency (MAF) was thresholded

at > 0.1 in the populations being compared for all comparisons except when calculating distances between African Americans and each of the African populations. To reduce the SNP ascertainment biases associated with SNP discovery in the YRI, we used only markers with a MAF > 0.1 in Europeans for the F_{ST} estimates.

Admixture analysis. Our local genomic PCA admixture method first normalizes the genotype matrix of all individuals using the procedure as in *eigenstrat* [Patterson et al., 2006]. Each chromosome is divided into 15 SNP non-overlapping windows. The score for an individual for a given window is the product of an individuals normalized and scaled genotypes across this window with the corresponding segment of the PC1 eigenvector (see Appendix C for more details of the procedure). Windows which have one or more missing genotypes for an individual are not given a score and are omitted by the Hidden Markov Model (HMM). This gives a vector of scores for each individual across all chromosomes. We assume that ancestral population scores are drawn from a normal distribution, and use the ancestral population sample means and variances as the estimated parameters for the distribution (see Appendix C for mathematical details of the model and validation).

Acknowledgements

We thank K. King for her work managing and preparing the POPRES data. We thank J. Degenhardt for helpful discussions and suggestions throughout the project, and K.E. Lohmueller for discussion, LD scripts and con-

structive comments on the manuscript. This work was supported by NIH (grant 1R01GM83606). SAT additionally acknowledges support by the NIH (grant R01GM076637), NSF (grants BCS-0196183, BSC-0552486, BCS-0827436) and David and Lucile Packard and Burroughs Wellcome Foundation Career Awards.

CHAPTER 4

GENOME-WIDE PATTERNS OF POPULATION STRUCTURE AND ADMIXTURE AMONG HISPANIC/LATINO POPULATIONS*

*Originally published as: K. Bryc, C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds, A. Auton, M. Hammer, C. D. Bustamante, and H. Ostrer (2010). *Proc Natl Acad Sci*, 107 Suppl 2:8954-61

4.1 Abstract

Hispanic/Latino populations possess a complex genetic structure that reflects recent admixture among and potentially ancient substructure within Native American, European, and West African source populations. Here, we quantify genome-wide patterns of SNP and haplotype variation among 100 individuals with ancestry from Ecuador, Colombia, Puerto Rico, and the Dominican Republic genotyped on the Illumina 650K platform and 112 Mexicans genotyped on Affymetrix 500K platform. Intersecting these data with previously collected high-density SNP data from 4,305 individuals, we use principal component analysis and clustering methods *FRAPPE* and *STRUCTURE* to investigate genome-wide patterns of African, European, and Native American population structure within and among Hispanic/Latino populations. Comparing autosomal, X and Y chromosome, and mtDNA variation, we find evidence of a significant sex bias in admixture proportions consistent with disproportionate contribution of European male and Native American female ancestry to present day populations. We also find that patterns of linkage-disequilibria in admixed Hispanic/Latino populations are largely impacted by the admixture dynamics of the populations with faster decay of LD in populations of higher African ancestry. Finally, using the locus-specific ancestry inference method *LAMP*, we reconstruct fine-scale chromosomal patterns of admixture. We document moderate power to differentiate among potential sub-continental source populations within the Native American, European, and African segments of the admixed Hispanic/Latino genomes. Our results suggest future genome-wide association scans in Hispanic/Latino populations may require correction for local genomic ancestry at a sub-continental scale when associating differences in the genome

with disease risk, progression and drug efficacy, as well as for admixture mapping.

4.2 Introduction

The term, Hispanic/Latinos, refers to the ethnically diverse inhabitants of Latin America and to people of Latin American descent throughout the world. Present-day Hispanic/Latino populations exhibit complex population structure with significant genetic contributions from Native American and European populations (primarily involving local indigenous populations and migrants from the Iberian peninsula and Southern Europe) as well as West Africans brought to the Americas through the trans-Atlantic slave trade [Sans, 2000, Wang et al., 2008]. These complex historical events have impacted patterns of genetic and genomic variation within and among present-day Hispanic/Latino populations in a heterogeneous fashion resulting in rich and varied ancestry within and among populations as well as marked differences in the contribution of European, Native American, and African ancestry to autosomal, X chromosome, and uniparentally inherited genomes.

Many key demographic variables differed among colonial Latin American populations, including the population size of the local pre-Columbian Native American population, the extent and rate at which European settlers displaced native populations, whether or not slavery was introduced in a given region, and, if so, the size and timing of introduction of the African slave populations. There were also strong differences in ancestry among social classes in colonial (and post-colonial) populations with European ancestry often cor-

relating with higher social standing. As a consequence, present day Hispanic/Latino populations exhibit very large variation in ancestry proportions (as estimated from genetic data) not only across geographic regions [Sans, 2000, Wang et al., 2008], but also within countries themselves [Seldin et al., 2007, Silva-Zolezzi et al., 2009]. In addition, the process of admixture was apparently sex-biased and preferentially occurred between European males and Amerindian and/or African females, and this process has been shown to be remarkably consistent among countries and populations including Argentina [Dipierri et al., 1998], Ecuador [González-Andrade et al., 2007], Mexico [Green et al., 2000], Cuba [Mendizabal et al., 2008], Brazil [Marrero et al., 2007], Uruguay [Sans et al., 2002], Colombia [Carvajal-Carmona et al., 2003], and Costa Rica [Carvajal-Carmona et al., 2003].

The rich diversity of variation in ancestry among Hispanic/Latino populations coupled with consistent differences among populations in the incidence of chronic heritable diseases suggests Hispanic/Latino populations may be very well-suited for admixture mapping [Smith et al., 2001, González Burchard et al., 2005]. For example, differences in relative European ancestry proportions correlate with higher susceptibility in Puerto Ricans to asthma as compared to Mexicans [Salari et al., 2005]. Data have also shown an increased risk of breast cancer in Latinas with greater European ancestry [Fejerman et al., 2008] and an interplay between African ancestry and cardiovascular disease and hypertension in Puerto Ricans from Boston [Lai et al., 2009]. Hispanic/Latinos are also likely to play an increasingly important role in multi- and trans-ethnic genetic studies of complex disease. Genome-wide scans have identified candidate markers for onset of type 2 diabetes in

Mexican-Americans from Texas [Hayes et al., 2007] as well as region on chromosome 5 associated with asthma in Puerto Ricans [Choudhry et al., 2008].

Quantifying the relative contributions of ancestry, environment (including socio-economic status), and ancestry by environment interaction to disease outcome in diverse Hispanic/Latino populations will also be critical to applying a genomic perspective to the practice of medicine in the U.S. and in Latin America. For example, whereas European ancestry was associated with increased asthma susceptibility in Puerto Ricans [Salari et al., 2005], it was also shown that the effect was moderated by socioeconomic status [Choudhry et al., 2006]. This suggests that quantifying fine-scale patterns of genomic diversity among diverse U.S. and non-U.S. Hispanic/Latino may be critical to the efficient and effective design of medical and population genomic studies. A fine-scale population genomics perspective may also provide a powerful means for understanding the roles of ancestry, genetics, and environmental covariates on disease onset and severity [González Burchard et al., 2005].

Here, we introduce a larger, high-density SNP and haplotype dataset to investigate historical population genetics questions - such as variation in sex biased ancestry and genome-wide admixture proportions within and among Latino populations - as well as provide a genomic resource for the study of population substructure within putative European, African, and Native American source populations. Our dataset includes three Latino populations that are underrepresented in whole-genome analyses, Dominicans, Colombians, and Ecuadorians, as well as Mexicans and Puerto Ricans–the two largest Hispanic/Latino ethnic groups in the U.S. This allows for the comparison of patterns of population structure and ancestry across multiple U.S.

Hispanic/Latino populations. Our dense SNP marker panel is formed by the intersection of two of the most commonly used genotyping platforms, allowing for the inclusion of dozens of Native American, African and European populations for ancestry inference. Our work expands on high-density population-wide genotype data from the International HapMap Project (HapMap) [Altshuler et al., 2005, Frazer et al., 2007], the Human Genome Diversity Panel (HGDP) [Rosenberg et al., 2002], and the Population Reference Sample (POPRES) [Nelson et al., 2008] that have representation of Mexicans, but not other Hispanic/Latino groups either from the Caribbean or from South America, with a resulting gap for analyzing admixture in those populations. This project, therefore, represents a first step towards comprehensive panels for U.S.-based studies that can more accurately reflect the diversity within various Hispanic/Latino populations.

4.3 Results

Population Structure We applied the clustering algorithm *FRAPPE* to investigate genetic structure among Hispanic/Latino individuals using a merged data set with over 5,000 individuals with European, African, and Native American ancestry genotyped across 73,901 SNPs common to the Affymetrix 500K array and the Illumina 650K genotyping panel (see Methods). *FRAPPE* implements a maximum likelihood method to infer the genetic ancestry of each individual, where the individuals are assumed to have originated from K ancestral clusters [Tang et al., 2005]. The plots for $K = 3$ and $K = 7$ are shown in Figure 4.1 and for all other values of K in Figure D.2. At K = 3 we observed clustering largely by Native American, African, and European ancestry, with the Hispanic/Latino

populations showing genetic similarity with all of these populations. However, significant population differences exist, with the Dominicans and Puerto Ricans showing the highest levels of African ancestry (41.8% and 23.6% African, SDs 16% and 12%), whereas Mexicans and Ecuadorians show the lowest levels of African ancestry (5.6% and 7.3% African, SDs 2% and 5%) and the highest Native American ancestries (50.1% and 38.8% Native American, SDs 13% and 10%). We also found extensive variation in European, Native American and African ancestry among individuals within each population. A clear example could be observed in the Mexican sample where ancestry proportions ranged from predominantly Native American to predominantly European (with generally low levels of African ancestry). Similar results were found in Colombians and Ecuadorians, whereas Dominicans and Puerto Ricans showed the greatest variation in the African ancestry (see Figure 4.1). Interestingly, at $K = 7$ we were able to capture signals of continental substructure such as a Southwest to Northeast gradient in Europe and a Native American component that is absent in the two Amazonian indigenous populations (Karitiana and Surui) but that substantially contributes to all other studied Latino populations. We also note that several of the individuals from the Maya and Quechua Native American samples (and to a lesser extent Nahua and Pima) from the Human Genome Diversity Panel (CEPH-HGDP) show moderate levels of European admixture, consistent with previous studies of these populations [Jakobsson et al., 2008]. Interestingly, this is not the case for the Aymara and Quechua samples genotyped by Mao et al. [Mao et al., 2007].

We also undertook principal component analysis (PCA) of the autosomal genotype data from Hispanic/Latino and putative ancestral populations using the smartpca program from the software package, *eigenstrat* (Figure 4.2A)

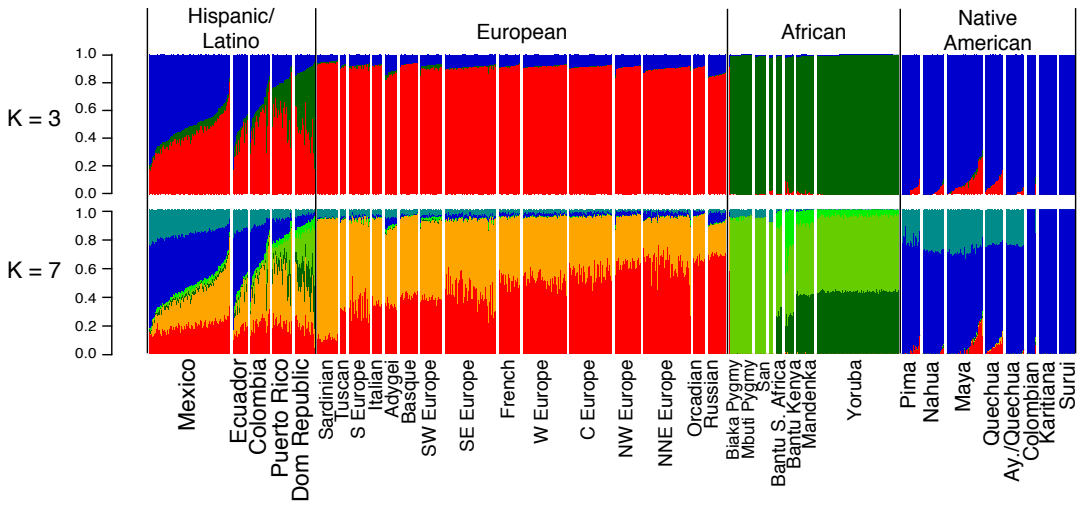

Figure 4.1: *FRAPPE* clustering illustrating the admixed ancestry of Hispanic/Latinos shown for $K = 3$ and $K = 7$. Individuals are shown as vertical bars colored in proportion to their estimated ancestry within each cluster. Native American populations are listed in order geographically, from North to South.

[Patterson et al., 2006]. The first two principal components of the PCA strongly support the notion that the three ancestral populations contributing to the Hispanic/Latino genomic diversity correspond exactly to Native American, European, and African ancestry. The Hispanic/Latino populations showed different profiles of ancestry, as exemplified by the fitting of ellipses to the covariance matrix of each populations first two PCs (Figure 4.2C). Subsequent PCs showed substructure within Africa, Native Americans, and Europeans (Figure D.2). PCA on the X chromosome markers (Figure D.2B) showed a similar pattern, although since there are only 1,500 markers this PCA had greater variance, which is illustrated in the fitted ellipses as well (Figure D.2D).

We also ran the Bayesian clustering algorithm *STRUCTURE* in assignment mode [Falush et al., 2003], and used a training set of Europeans, Africans, and Native Americans to estimate ancestral allele frequencies and assess admixture proportions within and among the Hispanic/Latino populations. Using

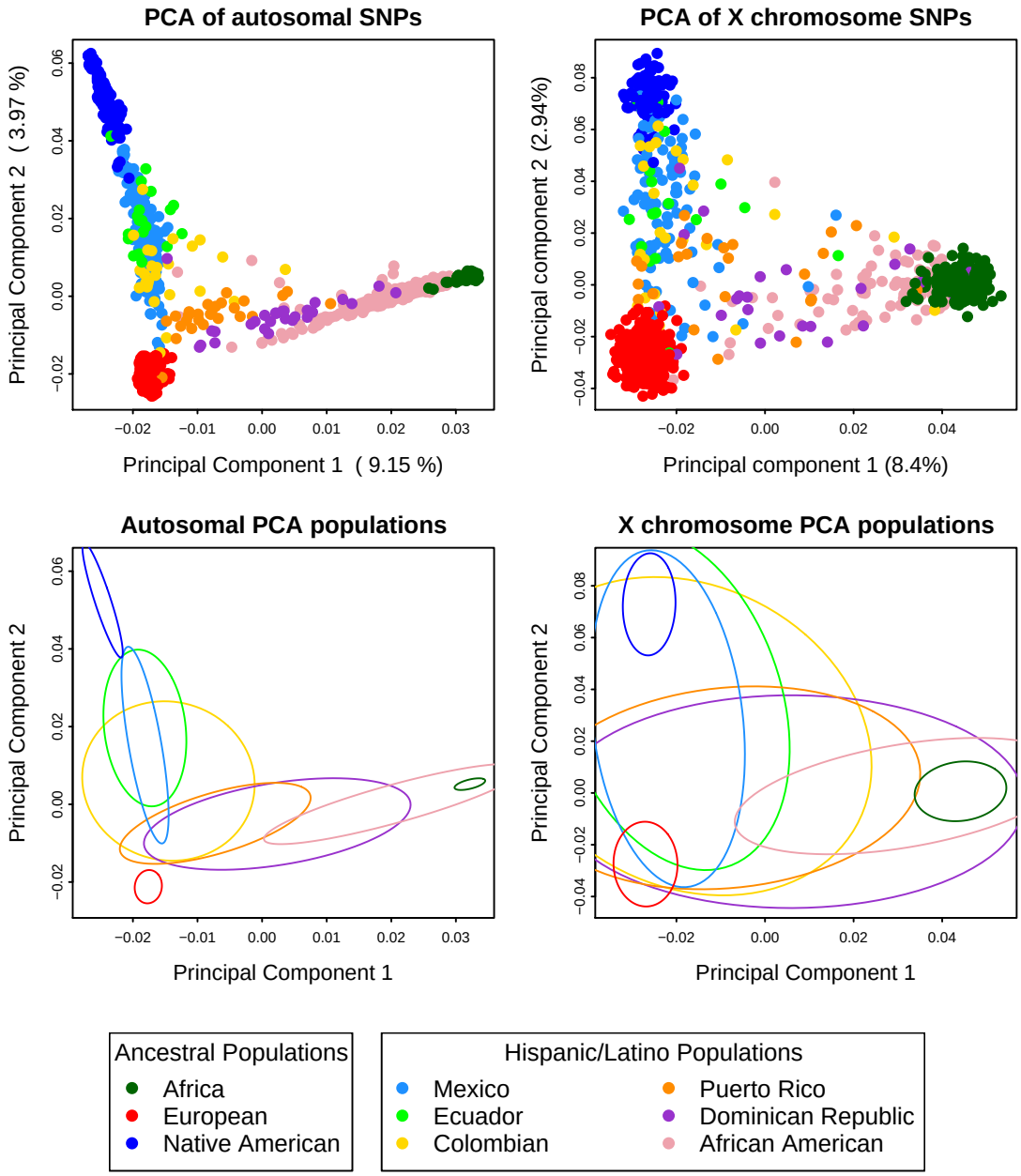

Figure 4.2: Principal component analysis results of the Hispanic/Latino individuals with Europeans, Africans, and Native Americans. PC 1 versus PC 2 scatter plots based on autosomal markers (top left) and based on X chromosome markers (top right). Ellipses are fitted to the PCA results on the autosomes (bottom left) and to results from the X chromosome markers (bottom right).

STRUCTURE analysis of the autosomes (Figure 4.3, top) and the X chromosome (Figure 4.3, bottom panel), we found that, again, Puerto Ricans and Dominicans showed the greatest proportion of African ancestry whereas Colombians, Ecuadorians and Mexicans showed extensive variation in European and Native American ancestry among individuals. We calculated LD decay curves for all populations with at least 10 individuals, choosing subsets of 10 individuals, and averaging over 100 random subsets of the data. Patterns of decay of LD were consistent with previously published results [Jakobsson et al., 2008] with Native American populations showing the highest levels of LD and African populations the lowest (Figure 4.4A). Interestingly, the Hispanic/Latino populations demonstrated rates of decay of LD that correlated strongly with the amount of Native American, European, and African ancestry (Figure 4.4B). Specifically, the populations with the most Native American ancestry, Mexican and Ecuadorian, exhibited higher levels of linkage disequilibrium among SNP markers, whereas the populations with the highest proportions of African ancestry, the Dominican and Puerto Rican samples, had the lowest levels of LD.

Locus-specific ancestry In order to reconstruct local genomic ancestry at a fine scale, we used the ancestry deconvolution algorithm, *LAMP*, [Sankararaman et al., 2008] allowing for a three-way admixture and focused on the four Hispanic/Latino populations genotyped on the Illumina 650K platformDominicans, Colombians, Puerto Ricans, and Ecuadorians (see Methods for details). Since this same SNP panel had also been genotyped across the HGDP samples (1,043 individuals from 53 populations), the merged data set containing more than 500,000 markers provided a unique resource for investigating the extent of subcontinental ancestry among diverse Hispanic/Latino populations.

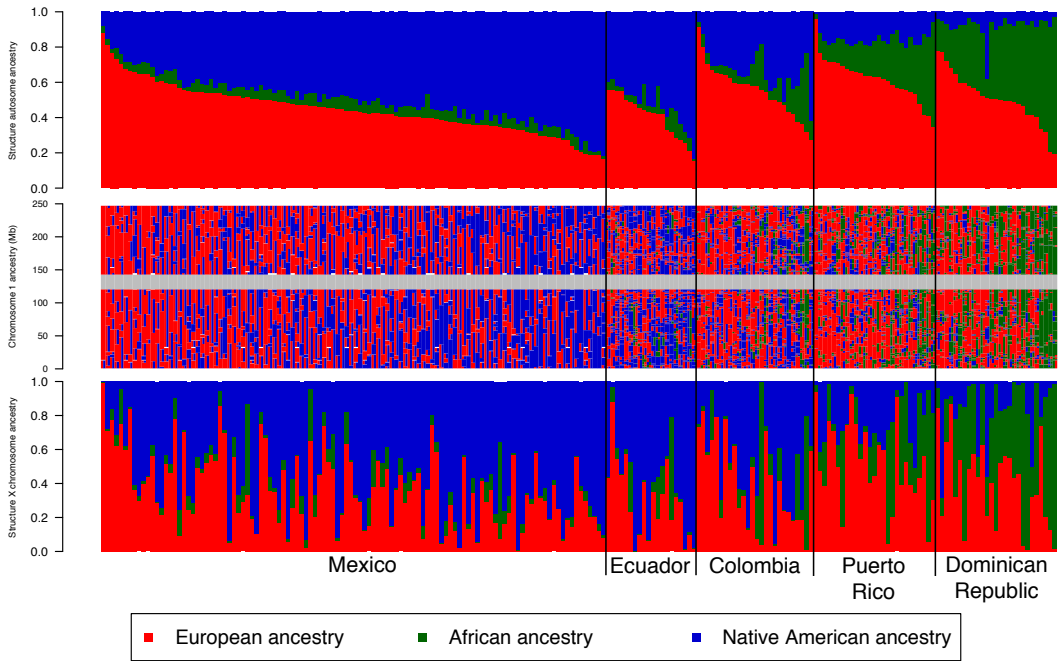

Figure 4.3: Genome-wide and locus specific ancestry estimates for Mexicans, Ecuadorians, Colombians, Puerto Ricans, and Dominicans. Shown for $K = 3$, *STRUCUTRE* clustering of the Hispanic/Latino individuals on the autosomes (top) and on the X chromosome (bottom). Individuals are shown as vertical bars colored in proportion to their estimated ancestry within each cluster. Local ancestry at each locus is shown for each individual on chromosome 1 (middle panel). The X chromosome shows greater Native American ancestry (blue) and greater variability in African ancestry (green), with reduced European ancestry (red).

We found that individual average ancestries are in agreement with *FRAPPE* and *STRUCTURE* results in which Ecuadorians have the highest Native American proportions, followed by Colombians (showing greater European contribution), and with Puerto Ricans and Dominicans showing the highest African ancestry – specially Dominicans who show very low contribution from Native Americans (see Figure 4.1). We also used the PCA-based methods of Bryc et al. [Bryc et al., 2010] to infer ancestry at each locus for the samples genotyped on the Affymetrix 500K which included over 100 Mexican samples genotyped by the POPRES project [Nelson et al., 2008] and diverse Native American popula-

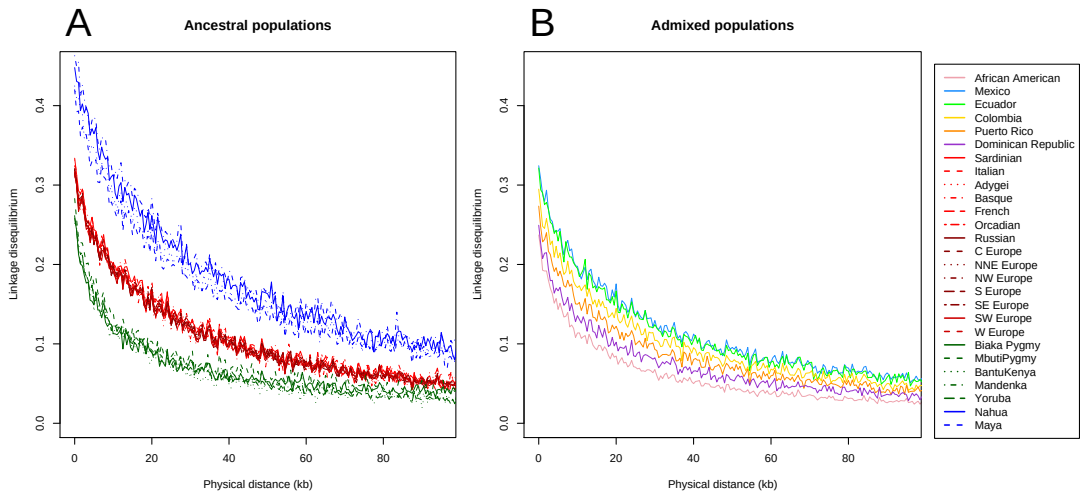

Figure 4.4: Linkage disequilibrium, genotype r^2 estimated by *PLINK*, by population as a function of physical distance (Mb). Native American, European, and African populations shown on the left panel, and the Hispanic/Latino populations shown on the right panel with the same scale.

tions genotyped by Mao et al. [Mao et al., 2007]. The local admixture tracks for each individual are in large agreement with the genome-wide average ancestry proportions (see Figure 4.3, middle panel).

To investigate the genetic relationships among admixed Hispanic/Latino populations and putative ancestral groups, we compare patterns of population divergence among the inferred segments of European, African, and Native American ancestry and corresponding putative source populations using Wright's F_{ST} measure. Specifically, we used *LAMP* to reconstruct for each individual in our data set, segments of European, African, and Native American ancestry across both the maximal SNP data set for all of the admixed and putative source population individuals (i.e., either the 650K Illumina for Puerto Rican, Ecuadorian, Columbian, and Dominican or 500K for Mexicans from Guadalajara) as well as approximately 70K SNPs common to both platforms. To calculate F_{ST} at a given SNP for a given pair of populations, we included only individuals

with unambiguous ancestry assignment (i.e., individuals with two European-, two Native American-, or two African-origin chromosomes). One potential confounder for this analysis is that sample sizes differ substantially among subpopulations within major continental regions (e.g., in the Native American set, we have sample sizes that range from $n = 7$ for Colombian indigenous Americans in HGDP to $n = 29$ for Nahua from Mexico in Mao et al. dataset). To minimize the potential bias of differences in sample size, we randomly selected $n = 7$ individuals from all potential subpopulations and recomputed Wright's F_{ST}. As seen in Table 4.1, we found that consistent with historical records, our results show that African segments of the Hispanic/Latino populations are more closely related to the Bantu-speaking populations of West Africa than other populations. Specifically, we found that the Colombians and Ecuadorians are most closely related to the Kenyan Bantu populations, whereas the Puerto Ricans and Dominicans are most close to the Yoruba from Nigeria. Likewise, European segments show the lowest F_{ST} values when compared to Southwest European populations (individuals from Spain and Portugal), as well as French and Italian individuals. Native American segments of the Hispanic/Latino individuals show the least genetic differentiation with Mesoamerican (e.g. Maya and Nahua), Chibchan (e.g. Colombian), and Andean (e.g. Quechua) populations. The closest relationship is clearly observed between Mexicans from Guadalajara and Nahua indigenous individuals.

Sex bias in ancestry contributions We used the *STRUCTURE* ancestry estimates on the autosomes and X chromosome to estimate Native American, European, and African, ancestry proportions of each Hispanic/Latino individual. We then compared the estimates of ancestry for each population on

Table 4.1: Ancestry-specific F_{ST} distances between Hispanic/Latino populations and different putative source populations

African segments of the genome	COL	DOM	ECU	PRI	
Bantu Kenya	3.19%	1.56%	6.10%	2.50%	
Bantu S. Africa	3.38%	1.48%	6.88%	2.54%	
Biaka Pygmy	6.52%	4.66%	10.14%	5.76%	
Mandenka	3.68%	1.42%	6.40%	2.38%	
Mbuti Pygmy	11.22%	8.88%	14.70%	10.22%	
YRI	3.26%	0.91%	6.48%	2.18%	
European segments of the genome	**COL**	**DOM**	**ECU**	**PRI**	**Mexico**
Adygei	1.84%	1.56%	1.67%	1.81%	1.01%
Basque	1.35%	1.13%	1.46%	1.53%	0.78%
EuropeanESE	1.39%	1.07%	1.23%	1.39%	0.56%
EuropeC	0.98%	0.69%	1.01%	1.06%	0.34%
EuropeNNe	1.25%	0.92%	1.21%	1.35%	0.44%
EuropeNW	1.24%	0.94%	1.10%	1.25%	0.44%
EuropeS	1.03%	0.71%	1.01%	1.11%	0.19%
EuropeSE	1.02%	0.78%	1.01%	1.18%	0.31%
EuropeSW	0.86%	0.54%	0.84%	0.92%	0.12%
EuropeW	1.08%	0.73%	1.10%	1.16%	0.27%
French	0.88%	0.61%	0.80%	0.94%	0.27%
Italian	0.89%	0.61%	0.85%	0.88%	0.27%
Orcadian	1.41%	1.09%	1.42%	1.51%	0.79%
Russian	1.65%	1.41%	1.37%	1.82%	0.88%
Sardinian	1.55%	1.27%	1.61%	1.57%	0.85%
Tuscan	1.05%	0.83%	0.93%	1.04%	0.34%
Native American segments of the genome	**COL**	**DOM**	**ECU**	**PRI**	**Mexico**
Aymara	4.01%	5.14%	4.24%	5.87%	2.40%
Colombian	5.30%	5.87%	5.80%	6.62%	4.19%
Karitiana	9.10%	9.06%	9.18%	10.12%	8.20%
Maya	4.72%	4.26%	5.45%	6.62%	1.42%
Nahua	3.61%	3.60%	4.15%	4.80%	0.57%
Pima	8.56%	9.31%	9.19%	10.58%	5.11%
Quechua	3.43%	3.15%	3.08%	5.17%	2.09%
Surui	13.80%	13.74%	13.77%	15.09%	11.06%

the autosomes versus on the X chromosome (Figure 4.5 and Figures D.3 and D.4). Whereas we found that Native American ancestry is significantly higher on the X chromosome than on the autosomes (including those populations with reduced Native American ancestry, i.e. Puerto Ricans and Dominicans), the autosomal versus X-chromosome difference was more attenuated with regards to African ancestry. This reduced deviation is present even in those Hispanic/Latino populations analyzed whose non-European ancestry was principally Native American in origin (i.e. Mexicans and Ecuadorians). Furthermore, greater Native American ancestry on the X chromosome in Puerto Ricans did not necessarily imply greater Amerindian ancestry on the autosomes. This finding is similar to those found by analyzing fine-scale genome pattern of popu-

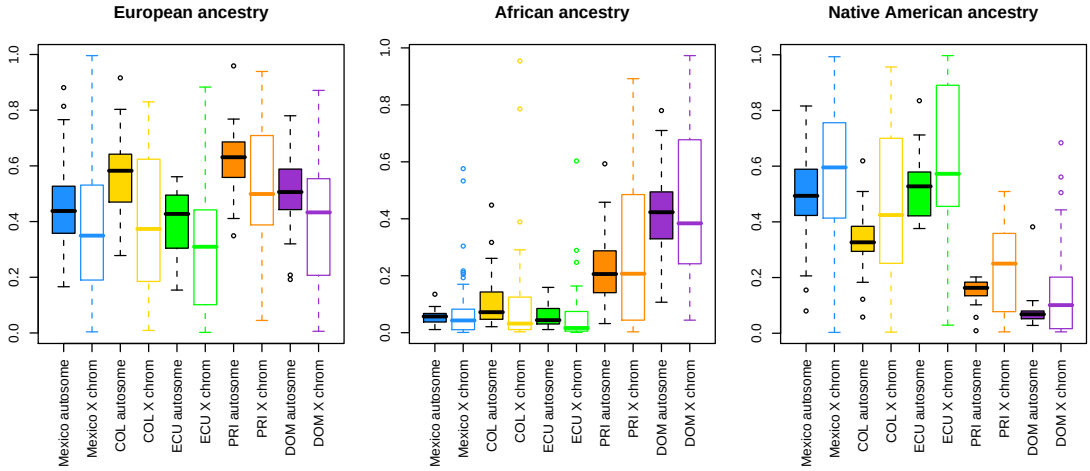

Figure 4.5: Boxplots comparing autosomal versus X chromosome ancestry proportions by population, shown for European ancestry (left), Native American ancestry (middle), and African ancestry (right). Filled boxes correspond to autosomal ancestry estimates while hollow boxes show X chromosome ancestry estimates. Median (solid line), first and third quartiles (box) and the minimum/maximum values, or to the smallest value within 1.5 times the IQR from the first quartile (whiskers). For each paired comparison of X chromosomes and autosomes, median Native American ancestries are consistently higher on the X chromosome in all Hispanic/Latino populations sampled, and European ancestries are lower across all populations.

lation structure and admixture among African Americans, West Africans, and Europeans [Lind et al., 2007].

Lastly, we used SNP and microsatellite genotyping to identify the canonical Y chromosome and mtDNA haplotypes for each of the Hispanic/Latino individuals we genotyped. We found an excess of European Y chromosome haplotypes and a higher proportion of Native American and African mtDNA haplotypes, consistent with previous studies (Figure 4.6). In addition, we found several non-European Y chromosomal haplotypes with most likely origins from North Africa and the Middle East. We observed that African-derived haplotypes were the predominant origin of mtDNA in Dominicans (17 out of 27 individuals), matching the greater African versus Native American ori-

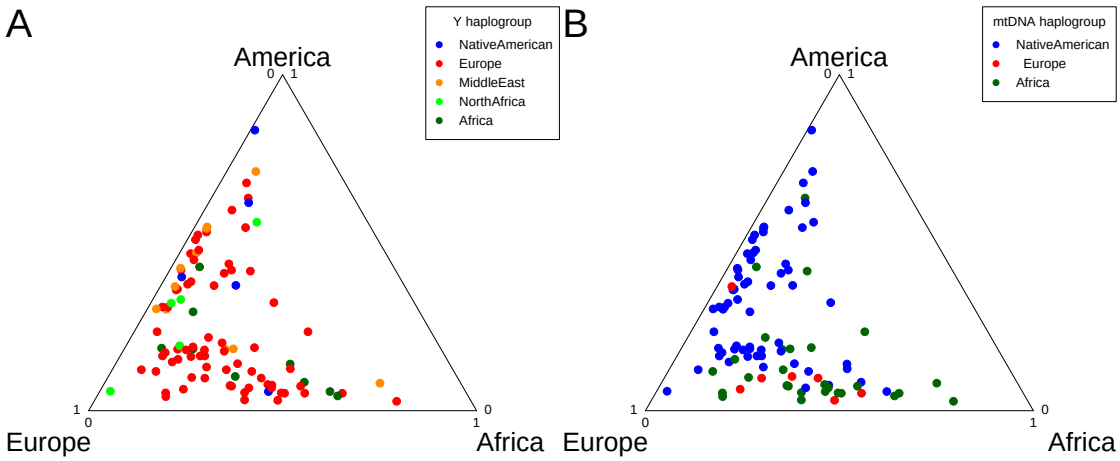

Figure 4.6: Comparison of mtDNA and Y chromosome haplotypes. Each individual is represented by a point within the triangle that represents the autosomal ancestry proportions. The most probable continental location for each individuals haplotype is designated by the color of the point. The Y chromosome contains a disproportionate number of European haplotypes, while the mtDNA has a high proportion of Native American, slightly more African haplotypes and fewer European haplotypes, consistent with a sex bias towards a great European male and Native American/African female ancestry in the Hispanic/Latinos.

gins of this population on the autosomes and X-chromosomes. However, in Puerto Ricans we did not find evidence of a high African female contribution. The predominant Y chromosomal origins in the Puerto Ricans sampled were European and African, but in contrast, 20 out of 27 Puerto Rican individuals had mitochondrial haplotypes of Native American origin, suggesting a strong female Native American and male European and African sex bias contribution. Overall, in all of the Hispanic/Latino populations we analyzed, we found evidence of greater European ancestry on the Y chromosome and higher Native American ancestry on the mtDNA and X chromosome consistent with previous findings [Dipierri et al., 1998, González-Andrade et al., 2007, Green et al., 2000, Mendizabal et al., 2008, Marrero et al., 2007, Sans et al., 2002, Carvajal-Carmona et al., 2003].

4.4 Discussion

Our work has important implications for understanding the population genetic history of Latin America as well as ancestry of US-based Hispanic/Latino populations. As has been previously documented, we found large variation in the proportions of European, African, and Native American ancestry among Mexicans, Puerto Ricans, Dominicans, Ecuadorians, and Colombians, but also within each of these groups. These trends are a consequence of variation in rates of migration from ancestral European and African source populations as well as population density Native Americans in pre-Columbian times [Sans, 2000]. We found that Dominicans and Puerto Ricans in our study showed the highest levels of African ancestry, consistent with historical records. European settlers to island nations in the Caribbean basin largely displaced Native American populations by the early to mid 16th century and concurrently imported large African slave populations for large-scale colonial agricultural production (largely of sugar). In contrast, Colombia has wider geographic differences ranging from Caribbean coasts to Andean valleys and mountains, which could explain the enrichment of African ancestry in some individuals and not in others, likely representing the differences in origin within Colombia. Finally, Mexico and Ecuador are two continental countries that had high densities of Native Americans during pre-Columbian times; as expected, the individuals from these two countries show the highest degree of Native American ancestry. Our findings clearly show that the involuntary migration of Africans through slave trade appears to have left a clear trace in Hispanic/Latino populations proximal to these routes.

From the F_{ST} analysis, we found that the high-density genotype data we

have collected is quite informative regarding the personal genetic ancestry of admixed Hispanic/Latino individuals. Specifically, we found that individuals differ dramatically within and among populations and that we can reliably identify subpopulations within major geographic regions (i.e., Europe, Africa, and the Americas) that exhibit lower pairwise F_{ST} (and, therefore, higher genetic similarity) to the inferred European, African, and Native American segments for the 212 individuals studied. We found, for example, that Nahua showed the lowest F_{ST} in Mexicans, consistent with the observation that the Nahua are one of the largest Native American populations in this region, and are likely to have contributed to the genomes of admixed individuals in Mexico (as opposed, for instance, to the Mexican Pima who fall outside the Mesoamerican cultural region and show considerably higher levels of differentiation). We also found that the lowest F_{ST} for the African regions of the Dominican and Puerto Rican genomes are with the Yoruba, a Bantu-speaking West African population that has been shown to be genetically similar to the African segments of African Americans sampled in the U.S. [Bryc et al., 2010]. Though we have limited Native American populations and Hispanic/Latino sample sizes and, thus, the differences in F_{ST} with different sub-continental populations suggest that there exists a reasonably strong signal of which present day populations are most closely related to the ancestral populations that contributed ancestry to each of the Hispanic/Latino populations.

When comparing inferred continental ancestry of the X and Y chromosomes and mitochondrial versus the autosomal genome, we observed an enrichment of European Y-chromosome versus autosomal genetic material, and a greater percentage of both Native American and African ancestry on the X-chromosomes and mtDNA compared to the autosomes for the Hispanic/Latino

individuals in this study. This suggests a predominance of European males and Native American/African females in the ancestral genetic pool of Latinos, consistent with previous studies. A particularly interesting observation from our work on sex-biased admixture is that the pattern exists not only within populations, but among Hispanic/Latino populations as well. In all populations studied, there is an enrichment of Native American ancestry both on the X chromosome and mtDNA compared to the autosomes. This would suggest that a greater female Native American contribution to the genome of Latinos. A different result was obtained in relation to African ancestry. We found a smaller difference between mean African ancestry on the X chromosome and the autosomes, compared to the difference in Native American ancestry. Furthermore, unlike in Native American ancestry, we found an overwhelming representation of Native American mtDNA haplogroups in Puerto Ricans, even though non-European ancestry on the autosomes was largely African.

It is important to note that this observation does not necessarily undermine the model of sex biased admixture among European males and African females in the founding of Hispanic/Latino populations, especially when one considers the predominance of European Y chromosomes in all groups studied. However, it suggests that admixture between European males and Amerindian/African females has been a complex process in the formation of the various Hispanic/Latino populations. Specifically, a reduced X versus autosome mean African ancestry compared to Native American ancestry suggests a more balanced gender contribution in the Hispanic/Latino genome by individuals of African ancestry. In the case of Puerto Ricans, the only way one can reconcile greater African ancestry on the X chromosome versus what would be expected on mitochondrial data would be through transmission of X chromosomes inde-

pendent of mitochondrial transmission, which is only plausible biologically via males. Caution, however, should be exercised before considering such conclusions as concrete; unlike X chromosomes which can recombine and, thus, represent haplotypes derived from thousands of individuals, mitochondrial DNA represents just one sole distant ancestor among these thousands. Thus, a larger mtDNA sample would be necessary compared to X chromosomes to have similar confidence that a cohort accurately reflects the presumed diversity of ancestry in the population as a whole.

The Y chromosomal results also demonstrate the insufficiency of the paradigm of European males and Native American/ African females to capture the complexity within the Latin American populations. For example, we find Y chromosomal haplotypes in Hispanic/Latinos with presume origin in the Middle East and Northern Africa. Given that historical documentation suggests that most of the non-African and non-Native American contribution to admixed Hispanic/Latino populations is from Southwest Europe, this suggests the contemporary populations inherited these Y chromosomes from Europeans, who, in turn, were descended from Middle Eastern or North African men. Several historical events could have led to the acquisition by Europeans of non-European haplotypes, perhaps during the period of the Roman Empire when the Mediterranean Sea behaved as a conduit (not a physical barrier) between Europe, the Middle East, and North Africa or by Sephardic Jews or Moorish Muslims during the European Middle Ages/Islamic Golden Age. Alternatively, the presence of non-European Y chromosomal haplotypes originating from the Middle East and North Africa could represent the result of Iberian Jews and Muslims (themselves admixed) fleeing the peninsula for New World territories in response to discriminatory policies that strongly pressured both communities

at the termination of the Reconquista. Essentially, the diversity of haplotypes in the Y chromosomes in Latinos reflects not only population dynamics from the fifteenth century onwards, but also the historical trends of population movement occurring across the Atlantic during centuries prior.

The marked genetic heterogeneity of Latino populations shown in this study, as previously suggested by other surveys of genetic ancestry [Wang et al., 2008, Mao et al., 2007, Price et al., 2007] has important implications for the identification of disease-associated variants that differ markedly in frequency among parental populations. In their study of 13 Mestizo populations from Latin America, for example, Wang et al. (2008) suggested that admixture mapping in Hispanic/Latino populations may be feasible within a two-population admixture framework since the mean African ancestry in Mestizo populations is typically low (< 10%) (2). Whereas this is true for Hispanic/Latino populations with origins in the continental landmass of the Americas (as the ones studied by Wang et al.), our results show that this may not apply for Latino populations with origins in the Caribbean as their African ancestry proportion is considerably higher and highly variable among individuals, suggesting an extensive three-way admixture and representing additional challenges for admixture mapping. Likewise, we find subtle but reproducible differences in subcontinental ancestry among Hispani/Latino individuals suggesting that even a three-way admixture model may not be sufficient to accurately model the dynamic population genetic history of these populations.

Another observation with important implications for designing association studies is the large variation in individual admixture estimates within certain Latino populations (e.g., Mexicans, Colombians and Ecuadorians). One

could expect such outcome when collecting samples from US-based Latino communities, which in turn may come from different locations within their countries of origin (e.g., Colombians and Ecuadorians). However, within the Mexican sample, which has been collected in a single sampling location (i.e., Guadalajara, Mexico), we also observe large variation in European vs. Native American admixture proportions. Our findings are in agreement with previous studies on genetic ancestry from Mexico City [Wang et al., 2008, Martinez-Marignac et al., 2007], supporting the idea that such urban agglomerations, where a large number of epidemiological studies are likely to take place, continue to host a wide range of genetic variability among individuals that may self-identify as individuals from the same population. Therefore, particular attention should be paid to carefully matching representative cases and controls, as well as to carefully control for ancestry when performing association studies using Hispanic/Latino populations. We hope our dense genome-wide admixture analysis has allowed greater insight into the population dynamics of multiple Hispanic/Latino populations, and provides a resource for designing next-generation epidemiological studies in these communities, opening the possibility of better understanding the genetic makeup of this growing segment of the U.S. population.

4.5 Materials and Methods

Datasets We genotyped 100 individuals with ancestry from Puerto Rico, the Dominican Republic, Ecuador and Colombia on Illumina 650K arrays. We extracted 400 European, 365 African American and 112 Mexican samples from the GlaxoSmithKline POPRES project, which is a resource of nearly 6,000 control in-

dividuals from North America, Europe, and Asia genotyped on the Affymetrix GeneChip 500K Array Set [Nelson et al., 2008]. We randomly sampled 15 individuals from each European country where possible, or the maximum number of individuals available otherwise, to select the POPRES European individuals to be included in our study. Further description of sampling locations, genotyping and data quality control are in reference [Nelson et al., 2008]. We include 165 and 167 individuals from the HapMap project from the CEU and YRI populations, thinned to the same SNP set [Frazer et al., 2007]. We also include all European, Native American, and African individuals from the HGDP genotyped on Illumina 650K arrays [Jakobsson et al., 2008]. Lastly, we include all Native American populations from the Mao et al. (2007) study genotyped on Affymetrix 500K arrays [Mao et al., 2007]. For each dataset, we used annotation information to determine the strand on which the data were given and to map all Affymetrix and Illumina marker ids to corresponding dbSNP reference ids (rsids). SNPs without valid rsids were excluded from analysis. Each dataset was then converted to the forward strand to facilitate merging of the data. Data from the various platforms were merged using the *PLINK* toolset, version 1.06 [Purcell et al., 2007]. Likewise, non-missing genotype calls that showed disagreement between datasets were omitted. All samples were approved by IRB protocols from their respective studies.

Data Quality Control The HapMap II release 23, HGDP, Mao et al. and POPRES samples were genotyped and called according to their respective quality control procedures [Frazer et al., 2007, Nelson et al., 2008, Jakobsson et al., 2008, Mao et al., 2007]. Our final merged dataset contains

73,901 SNPs with genotype missingness of < 0.1 and < 0.05 individual missingness across 5,104 individuals.

Population Structure We used the software *FRAPPE*, which implements an expectation-maximization algorithm for estimating individual membership in clusters [Tang et al., 2005]. This algorithm is more computationally efficient than other MCMC methods allowing it to analyze many more markers than, for example, *STRUCTURE* [Tang et al., 2005, Falush et al., 2003]. After thinning markers to have $r^2 < 0.5$ in 50 SNP windows, shifted and recalculated every 5 SNPs, we ran *FRAPPE* on all 64,935 remaining markers for 5,000 iterations. We also assessed admixture proportions for the Hispanic/Latino individuals using *STRUCTURE* on a reduced dataset of 5,440 markers after thinning for MAF> 0.2 and with a minimum separation of 400Kb between markers. We use the `F` model with `USEPOPINFO = 1` to update allele frequencies using only the ancestral individuals, with 5,000 burn-in and 5,000 iterations [Falush et al., 2003]. We also used all 1,518 SNPs on the X chromosome for the same analysis of the X chromosome ancestry. Principal component analysis was conducted using a dataset thinned to have $r^2 < 0.8$ in 50 SNP windows, leaving 69,212 SNPs for analysis using the package *smartpca* from the software *eigenstrat*. Ellipses were fitted following the means and one standard deviation of the variance-covariance matrix of the PC1 and PC2 scores of each population.

For local ancestry estimation, we used the software *LAMP* in *LAMPANC* mode providing allele frequencies for the HGDP West Africans, Europeans and Native Americans as ancestral populations [Sankararaman et al., 2008]. A total of 552,025 SNPs were included in the analysis and configuration parameters were set as follows: mixture proportions (alpha) = 0.2, 0.4, 0.4; number of gener-

ations since admixture (g) = 20; recombination rate (r) = 1e-8; fraction of overlap between adjacent windows (offset) = 0.2; and r^2 threshold (ldcutoff) = 0.1. Local ancestry estimation for the Mexican individuals was performed using the two-way PCA-based method described in Bryc et al [Bryc et al., 2010] for both the full Illumina 650K and the Affymetrix 500K datasets, in 10 SNP windows. Only Native Americans with < 0.01 European ancestry (as estimated from *FRAPPE* results) were used as the ancestral Native American individuals within their respective datasets. F_{ST} was calculated between Native American, European, and African regions of the Hispanic//Latino individuals and the respective continental populations using a C++ implementation of Weir and Cockerham's F_{ST} weighed equations from [Weir and Cockerham, 1984]. To eliminate bias in estimation of F_{ST} due to European ancestry shown in some of the Native Americans, we also removed regions showing European ancestry within any of the Native Americans showing ¿ 0.01 European ancestry, using the same local ancestry estimation procedure as described for the Mexican individuals. Furthermore, to avoid any potentially confounding effect of sample size, we used a random sample of 7 (the minimum sample size of the Native American populations) individuals per non-Hispanic/Latino population to calculate pairwise F_{ST}. MAF was set at a threshold > 0.1 in the populations compared by F_{ST} calculations.

Acknowledgements We thank Mariano Rey for his support of the project, Peter Gregersen, Carole Oddoux and Annette Lee for their technical assistance, and Marc Pybus for valuable programming support during part of the analyses. This work was supported by the National Institutes of Health (Grant

1R01GM83606) as part of the National Institute of General Medical Sciences research funding programs.

APPENDIX A

SUPPLEMENTAL INFORMATION FOR CHAPTER 1

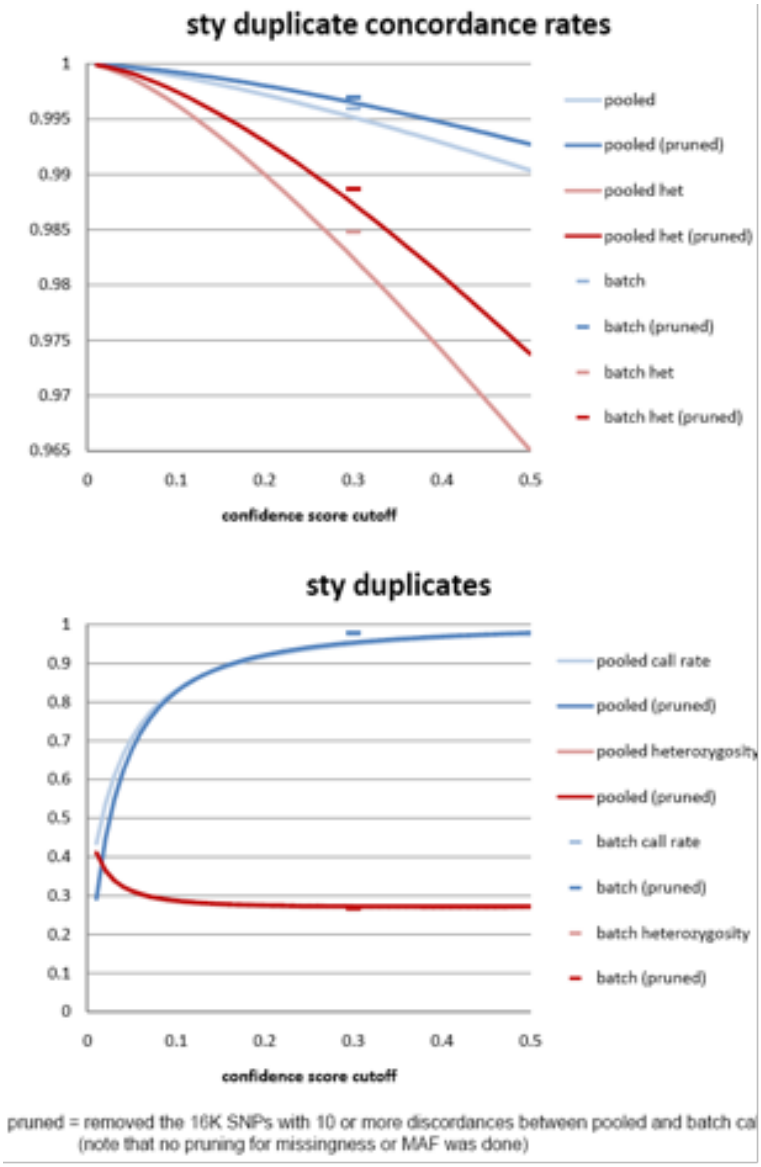

Figure A.1: Comparison of duplicate concordance and per subject call rates across BRLMM quality thresholds with the StyI chip. Concordance rates were determined for genotypes called simultaneously across all chips (pooled) or within groups of 48-96 chips genotyped at the same time (batch). Batch concordance rates were determined at the 0.3 confidence score cutoff, only. Results are presented for all genotypes and for heterozygous call only (het).

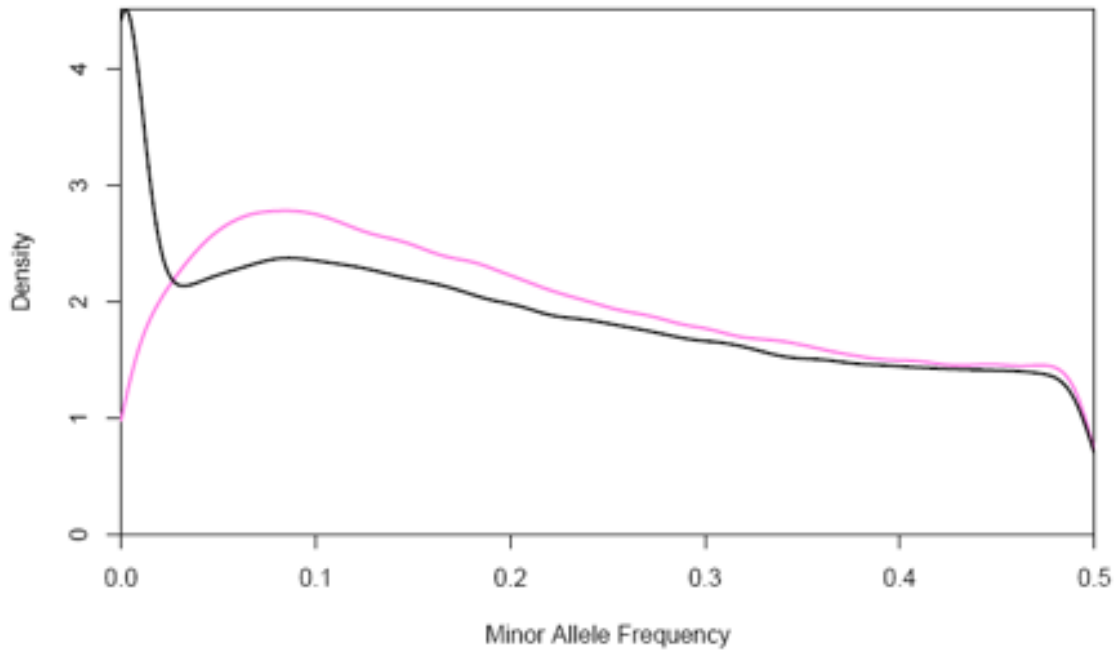

Figure A.2: Distribution of minor allele frequencies in POPRES African Americans (magenta) and HapMap Africans (black).

Table A.1: Country of origin of 21 abacavir-associated hypersensitivity reaction cases and the country-based matches selected.

Sex	Mother Country	Father Country	Control Match 1	Control Match 2
Male	Canada	Ukraine	10 Canada	
Female	Italy	Italy	10 Italy	
Male	Italy	Italy	10 Italy	
Male	Italy	Italy	10 Italy	
Male	Italy	Italy	10 Italy	
Male	Italy	Italy	10 Italy	
Male	Italy	Italy	10 Italy	
Male	Portugal	Portugal	10 Portugal	
Male	Portugal	Portugal	10 Portugal	
Male	Scotland	United Kingdom	2 Scotland	8 United Kingdom
Male	Spain	France	5 Spain	5 France
Male	Spain	Spain	10 Spain	
Male	Spain	Spain	10 Spain	
Male	Spain	Spain	10 Spain	
Male	Spain	Spain	10 Spain	
Female	United Kingdom	Netherlands	8 United Kingdom	2 Netherlands
Male	United Kingdom	United Kingdom	10 United Kingdom	
Male	USA	USA	5 USA	5 Canada
Male	USA	USA	4 USA	6 Canada
Male	USA	USA	4 USA	6 Canada
Male			10 Switzerland	

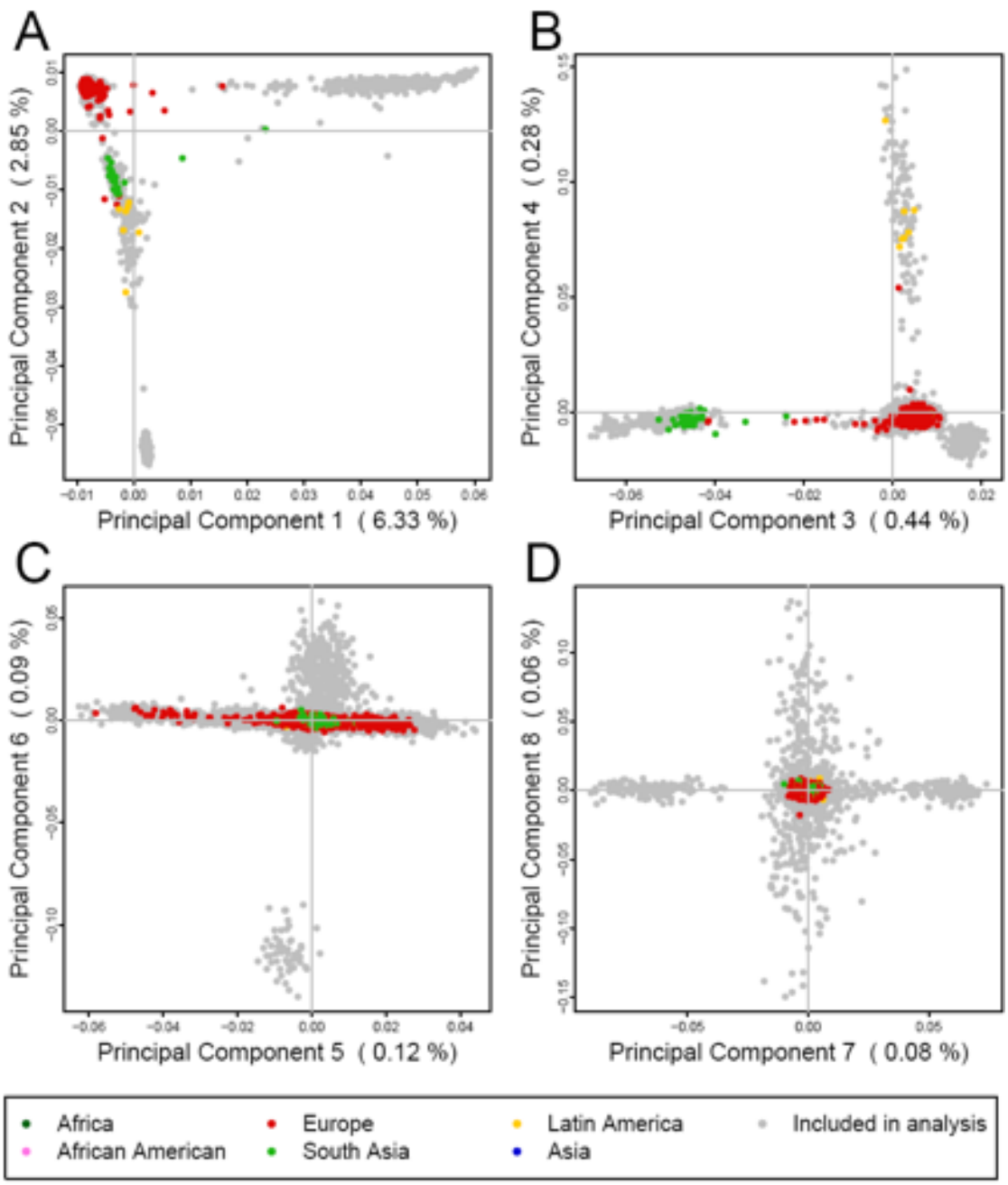

Figure A.3: Principal component scores for subjects that passed genotype quality control but were not included in the primary PCA. Scores were computed from factor loadings produced from the primary analysis. Scores from subjects included in primary analysis shown in gray.

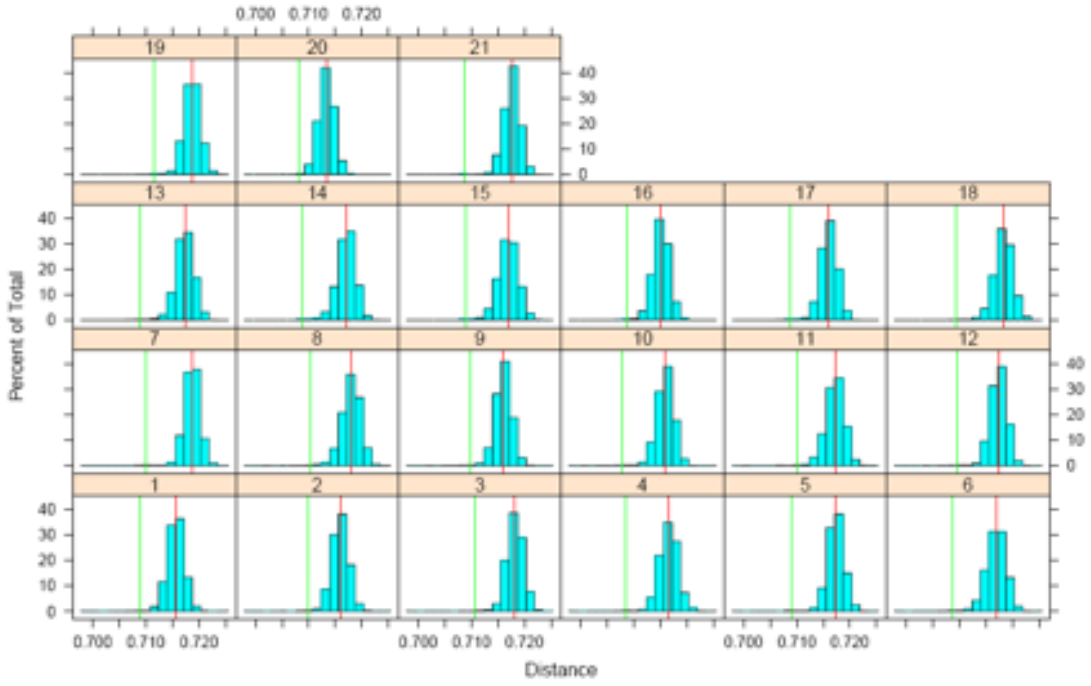

Figure A.4: Distribution of identity-by-state distance between each case and all POPRES subjects of European origin. Each panel shows the distribution of distances for each case. Vertical red line indicates the median distance. Vertical green line to the left indicates the position of the tenth closest subject used for matching controls to each case.

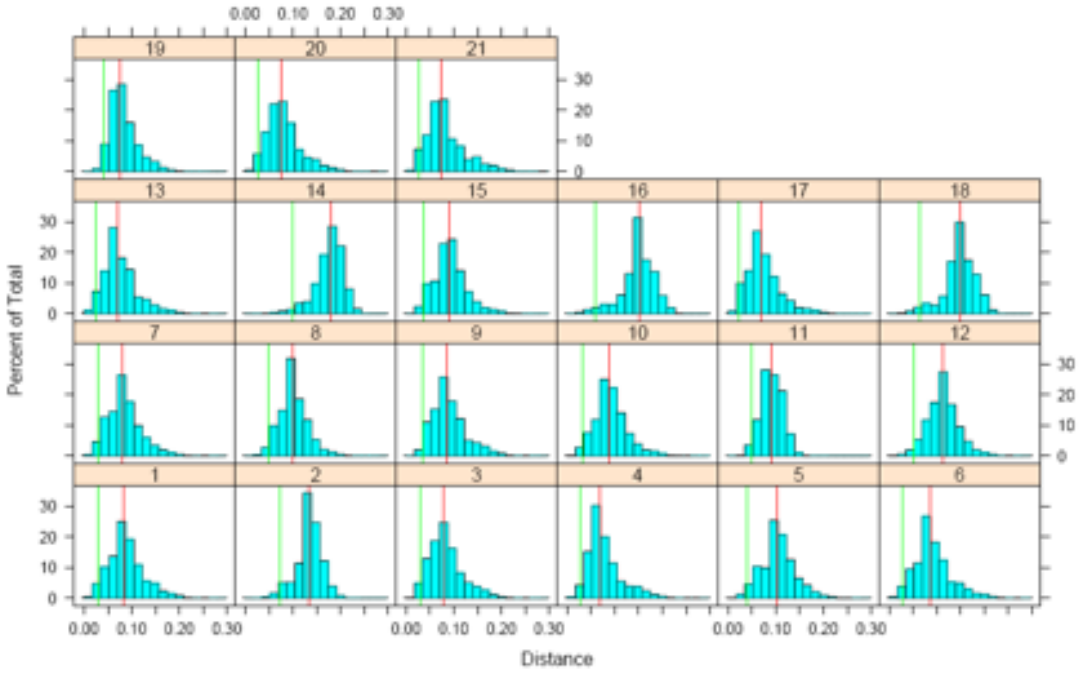

Figure A.5: Distribution of Euclidean distances from each case (panel) to each European POPRES subject on based first four principal components. Vertical red line indicates the median distance. Vertical green line to the left indicates the position of the 2.5th percentile from which matching controls were sampled.

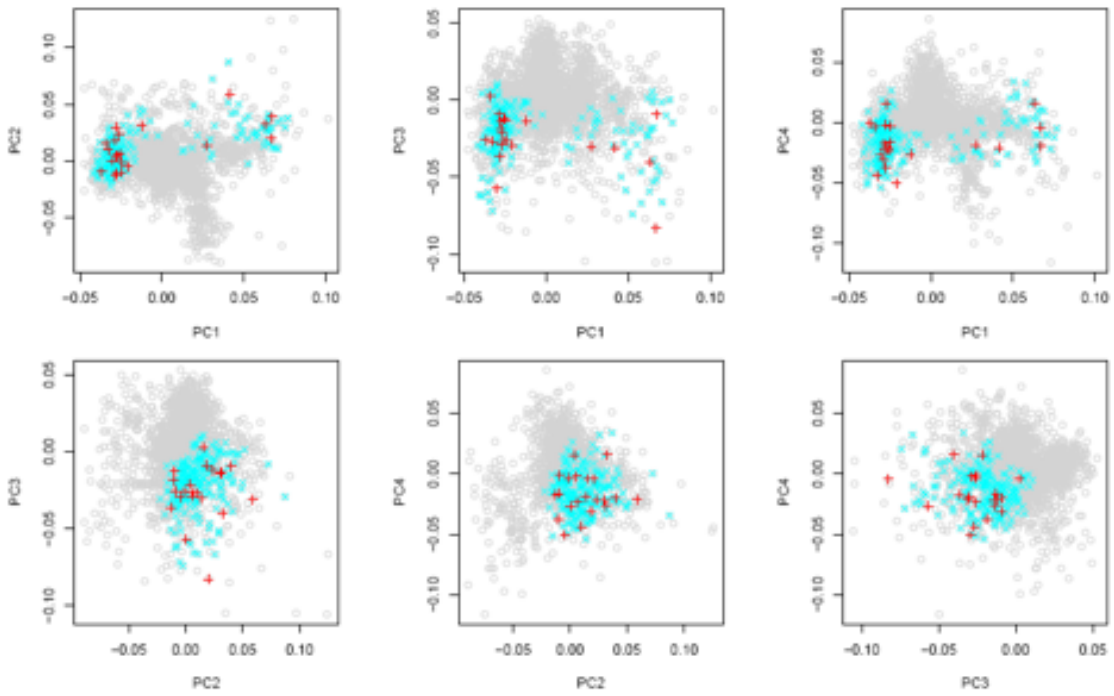

Figure A.6: Selection of ten controls to each case by use of principal component analysis. All cases and controls are plotted in each scatterplot based on their first four principal component scores. Cases are indicated by red +, controls selected on minimizing distance are indicated by blue x, and the remaining POPRES subjects by gray o.

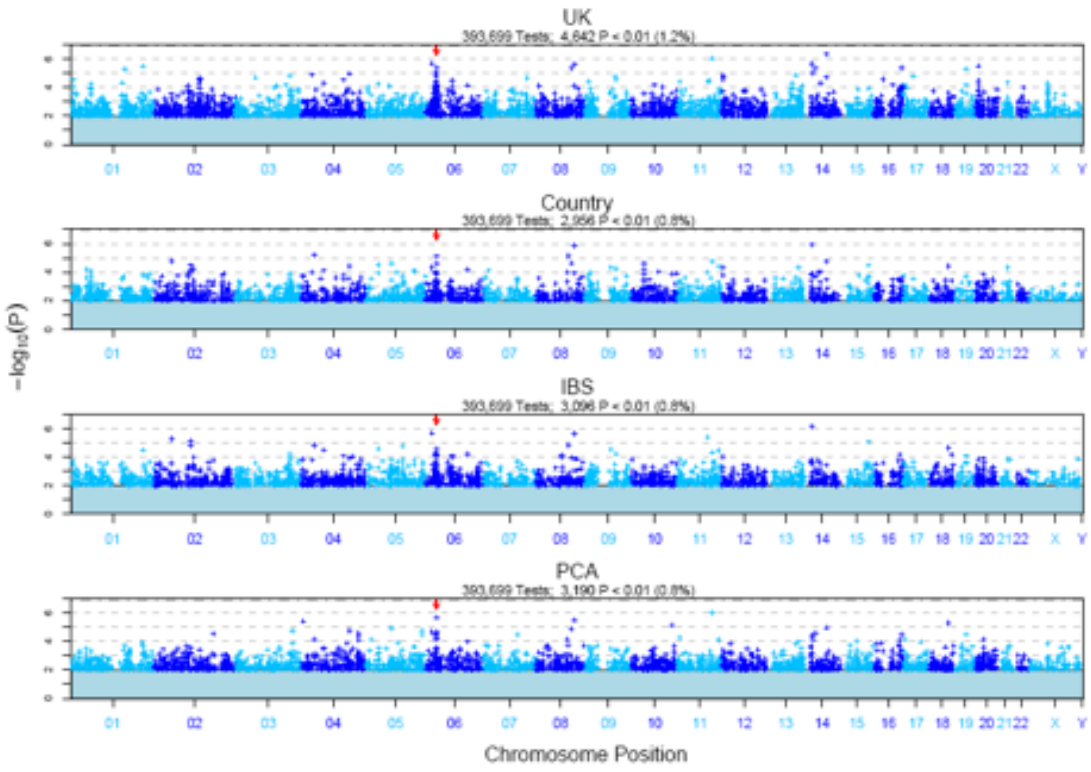

Figure A.7: Genome-wide plots of statistical significance of allelic tests (y axis) versus chromosome position (x axis) for Abacavir-associated hypersensitivity reaction pharmacogenetic case study. Controls are matched by European continent (UK), country, and minimizing identity by state (IBS) or principal component analysis score (PCA) distances. The position of the HLA-B gene is indicated by a red arrow.

APPENDIX B

SUPPLEMENTAL INFORMATION FOR CHAPTER 2

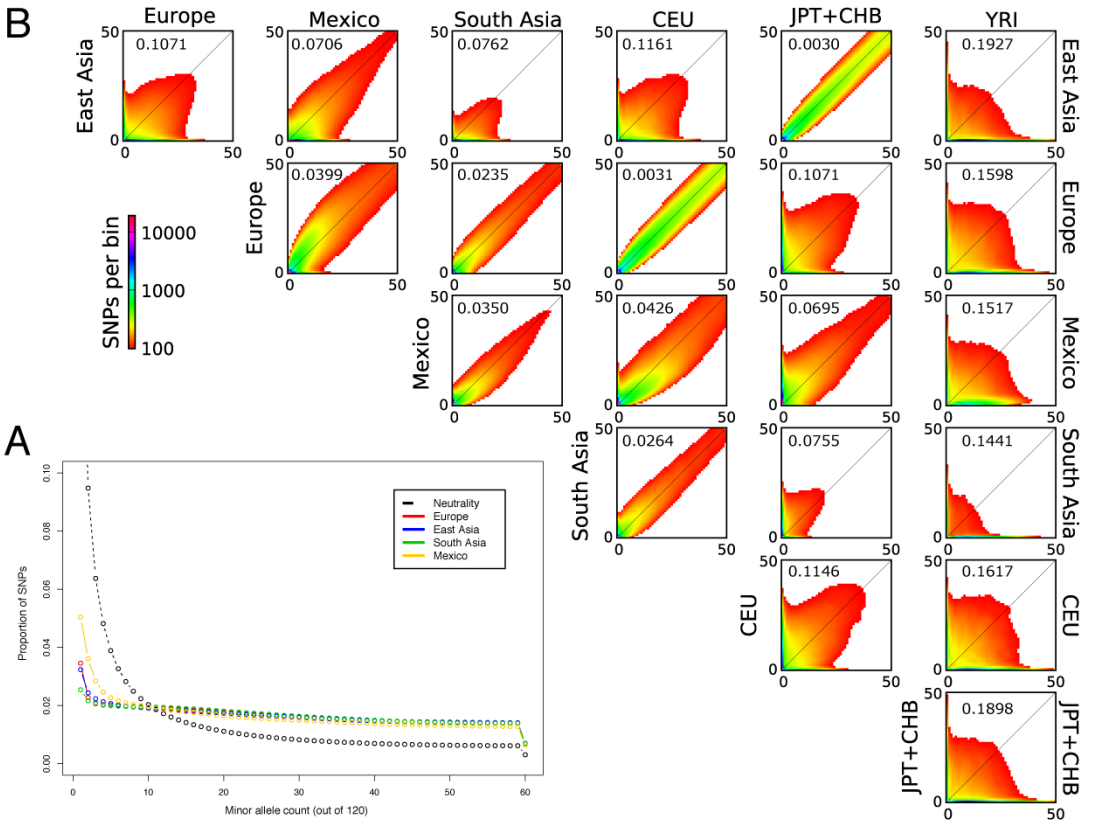

Figure B.1: Frequency spectra of the POPRES populations. (A) Minor Allele Frequency Spectra for the four sub-continental populations. The spectrum expected under neutrality is also shown in black. To account for differences in sample size, each sample was projected down to 120 chromosomes using the hypergeometric distribution. (B) Two-dimensional joint frequency spectra for each pairwise sub-continental population comparison. In this case, each sample was projected down to 100 chromosomes using a hypergeometric distribution. For each plot, the minor allele is defined from the total frequency in the two populations. Colors represent the number of SNPs within each bin. Entries in the spectra containing less than 100 SNPs are shown in white. Autosomal estimates of F_{ST} for each comparison are shown in the upper left hand corner of each figure.

Minor Allele Frequency Spectrum

The Affymetrix GeneChip provides a non-random sample of SNPs in the genome, with SNPs selected based on the catalog of known variants, frequency,

Table B.1: F_{ST} estimates between pairs of populations. Autosomal estimates are shown in the upper matrix triangle, whereas X chromosome estimates are shown in the lower triangle. For comparison, F_{ST} estimates using the HapMap populations and the same SNP set are also shown.

	East Asia	Europe	Mexico	South Asia	CEU	JPT+CHB	YRI
East Asia	-	0.1071	0.0706	0.0762	0.1161	0.0030	0.1927
Europe	0.1595	-	0.0399	0.0235	0.0031	0.1071	0.1598
Mexico	0.0965	0.0826	-	0.0350	0.0426	0.0695	0.1517
South Asia	0.1027	0.0426	0.0592	-	0.0264	0.0755	0.1441
CEU	0.1717	0.0056	0.0849	0.0456	-	0.1146	0.1617
JPT+CHB	0.0047	0.1560	0.0912	0.0987	0.1655	-	0.1898
YRI	0.3063	0.2640	0.2406	0.2300	0.2529	0.2928	-

and assay design considerations. The observed minor allele frequency (MAF; Supplementary Figure B.1A) spectrum is therefore not representative of the underlying true population allele frequency distribution. Nonetheless, patterns of correlated allele frequencies among populations (which largely reflect the history of divergence and migration between populations) can provide novel insights into average genealogical relationships among individuals from different populations (Supplementary Figure B.1B). From comparing the joint site-frequency spectra of common variation, we find SNP frequencies are more strongly correlated between Europe and South Asia than between East and South Asia. This result is consistent with a more severe founding bottleneck in the history of East Asian populations as well as less gene flow between South and East Asia than between South Asia and Europe.

Estimation of F_{ST}

F_{ST} was calculated using the 'strict' individuals from each population. We estimated F_{ST} for each SNP using the method of Weir andCockerham [Weir and Cockerham, 1984]. Specifically, we use equation 6 in that paper, and

for clarity, we repeat the formula here:

$$\hat{F}_{ST} = \frac{s^2 - \frac{1}{\bar{n}-1}\left[\bar{p}\,(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{\bar{h}}{4}\right]}{\left[1 - \frac{\bar{n}C^2}{(\bar{n}-1)r}\right]\bar{p}\,(1-\bar{p}) + \left[1 + \frac{\bar{n}(r-1)C^2}{(\bar{n}-1)r}\right]\frac{s^2}{r} + \left[\frac{C^2}{(\bar{n}-1)r}\right]\frac{\bar{h}}{4}}$$

where s^2 is the sample variance of allele frequencies over populations, $\bar{n}$ is the mean sample size, $\bar{p}$ is the mean sample allele frequency, r is the number of subpopulations, $\bar{h}$ is the mean heterozygote frequency in the sample, and C^2 is the squared coefficient of variation of the sample sizes. Further details are given in the cited paper. X chromosome estimates were obtained using only the female individuals in the study. To obtain a single estimate of F_{ST} for the complete data set, we combined estimates from all SNPs with a defined F_{ST} estimate using the weighted average scheme described in same paper (c.f. equation 10 in [Weir and Cockerham, 1984]).

To estimate the expected value of F_{ST} for the X chromosome based on autosomal F_{ST}, we use a standard result from population genetics that for an idealized Wright-Fisher population with migration among manydemes, the expected value of F_{ST} is simply:

$$E(F_{ST}) = \frac{1}{1 + 4Nm}$$

wheren $2Nm$ is the number of migrants entering each deme every generation. Under this condition, one can invert the expression to estimate Nm for the autosomes as $\widehat{Nm} = \frac{1}{4}\frac{1-F_{ST}}{F_{ST}}$. Under equal migration of males and females, equal variance in offspring number, and equal population size of the two sexes, the expected value for the X chromosome based on autosomal F_{ST} is then $\frac{1}{1+3\widehat{Nm}}$.

Subcontinental Population Structure Analysis

With the exception of Europe, sub-continental population structure analyses are described below. European population structure in the POPRES has been discussed elsewhere [Novembre et al., 2008], with the large sample size allowing population structure to be observed at a fine-scale. However, the POPRES provides evidence of structure in the other continental populations, even with their smaller sample sizes. In the following section, we describe patterns of population structure at a subcontinental level using both *STRUCTURE* and Principal Component Analysis (PCA). Note that markers were selected independently in each of the following analyses.

East Asia: For East Asia, we analyzed the POPRES individuals combined with the Han Chinese (CHB) and Japanese (JPT) samples from the HapMap. Using the subset of 271 individuals from East Asia, we ran *STRUCTURE* on 6,422 randomly selected SNPs with MAF > 0.2 (within East Asia) spaced 400kb apart. The results are shown in Supplementary Figure B.2B. As expected, at $K = 2$ we see two clear clusters separating the Japanese populations from the Chinese. At $K = 3$ we see that sections of the two different HapMap populations cluster together, reducing the proportion of genomes differentiated between Japanese and Chinese individuals. Further increasing K increases substructure within our POPRES samples not corresponding to known geographic structure.

In the PCA of the East Asian populations, we see clear separation between the Japanese and Taiwanese/Chinese samples (Figure 2.1C), with PC 1 separating the Japanese samples from Taiwan and the CHB — a pattern also seen in the *STRUCTURE* analysis. The second PC separates Taiwan from the HapMap Han Chinese, reflecting the geographic distance between these populations. To

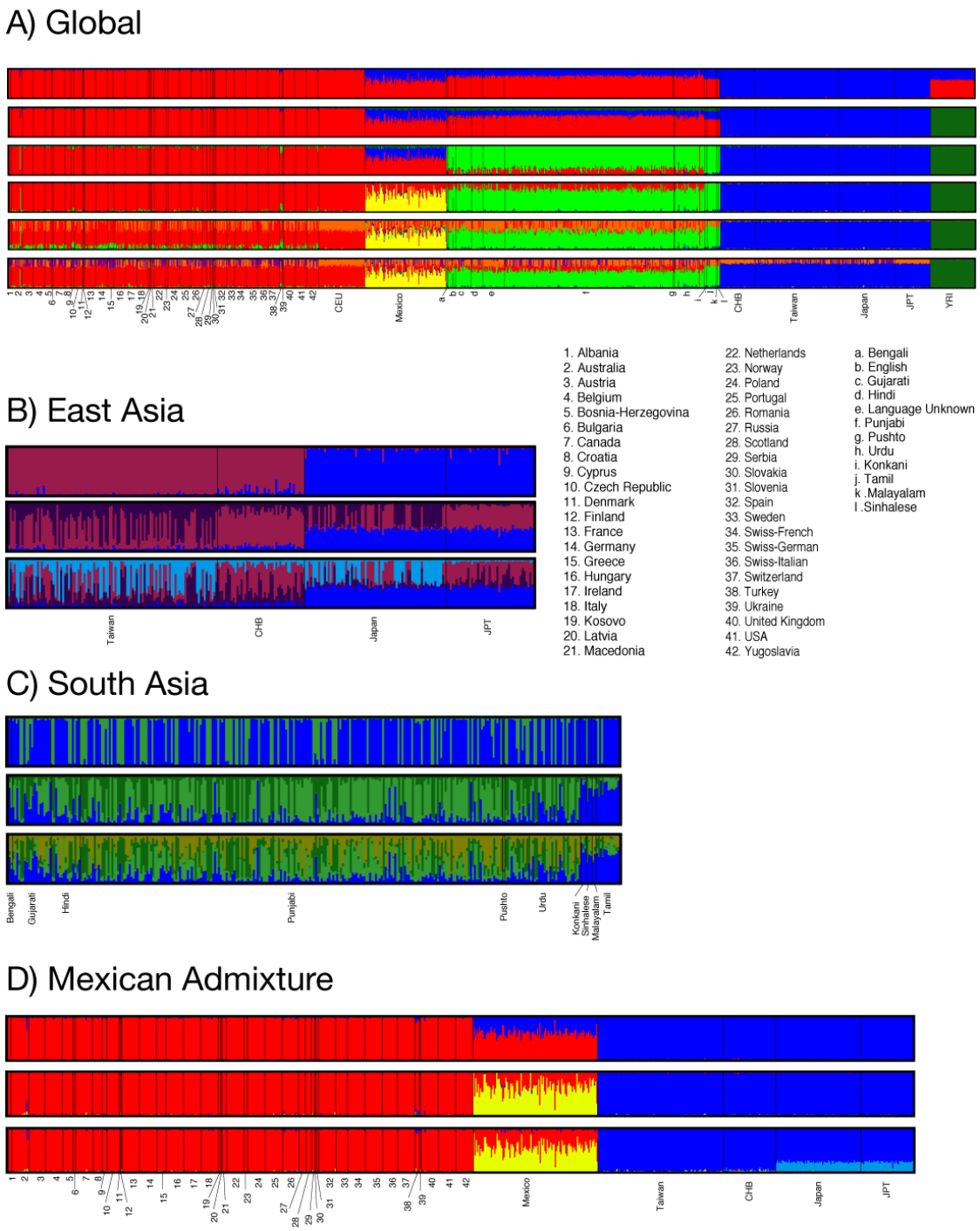

Figure B.2: (A) Global *STRUCTURE* results for K=2 to K=6. Subsequent plots show regional analyses for K=2 to K=4. (B) East Asia. (C) South Asia. (D) Mexican Admixture.

Table B.2: Details of population groupings.

Population Group	Continental Group	Individuals	'Strict' Individuals	Included Countries / Language Groups
Dravidian Influenced	South Asia	20	20	Konkani, Malayalam, Sinhalese, Tamil
Non-Dravidian Influenced	South Asia	312	284	Bengali, Gujarati, Hindi, Punjabi, Pushto, Urdu
Europe (C)	Europe	190	186	Austria, Germany, Netherlands, Switzerland (German)
Europe (ESE)	Europe	10	8	Cyprus, Turkey
Europe (NNE)	Europe	78	76	Czech Republic, Denmark, Finland, Hungary, Latvia, Norway, Poland, Russia, Slovakia, Sweden, Ukraine
Europe (NW)	Europe	459	447	Ireland, UK
Europe (S)	Europe	238	232	Italy, Switzerland (Italian)
Europe (SE)	Europe	99	96	Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Kosovo, Macedonia, Romania, Serbia, Slovenia
Europe (SW)	Europe	272	264	Portugal, Spain
Europe (W)	Europe	1,063	1,042	Belgium, France, Switzerland (French)
Mexico	Central America	112	107	Mexico
Japan	East Asia	73	73	Japan
Taiwan	East Asia	108	108	Taiwan
Europe Other A	Europe	237	0	Apparent European ancestry, but self-identified from the USA, Canada or Australia
Europe Other B	Europe	18	0	Apparent European ancestry, but self-identified from elsewhere
Europe (Mixed)	-	524	0	European individuals of mixed ancestry
South Asian Other	South Asia	28	0	South Asian individuals without language information
Unknown	-	4	0	No geographic or linguistic information

a much lesser extent the second PC also separates the POPRES Japanese from the HapMap Japanese. We note that the HapMap individuals were sampled in Tokyo, Japan, whereas the POPRES Japanese were sampled in Sydney, Australia [Nelson et al., 2008]. In the absence of further ancestral information, it is difficult to assess whether the small observed separation between the Japanese samples is due to subtle genotyping platform differences or true genetic differences.

South Asia: South Asian individuals were sampled as part of the LOLIPOP study in London, England, and we do not have data regarding parental or grand-parental ancestry of these individuals. However, we do have information regarding self-identified country of origin and spoken language with 10

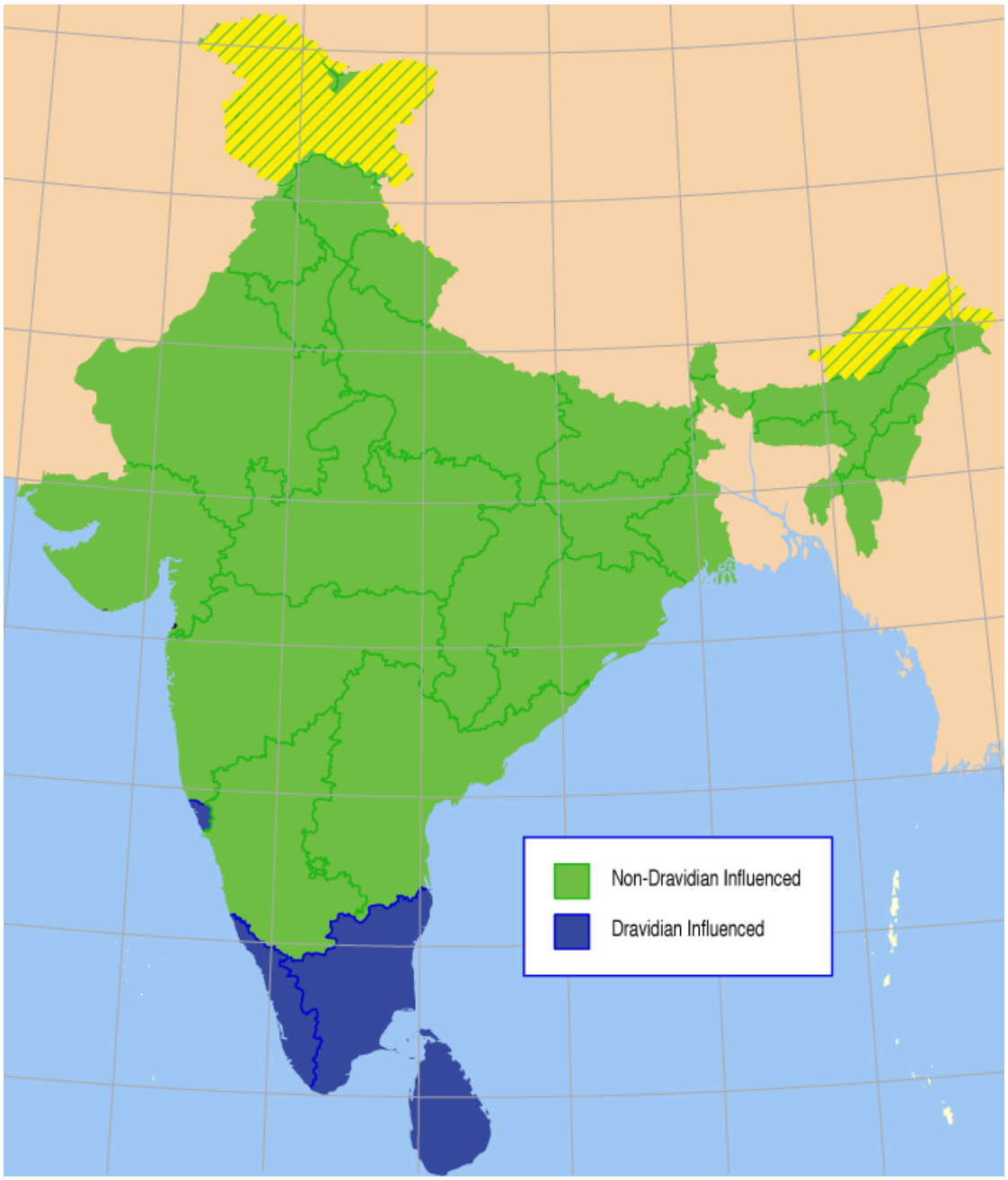

Figure B.3: Map of India and Sri Lanka showing the regions in which the Dravidian Influenced languages (blue) and the Non-Dravidian Influenced languages (green) are spoken, based on the official languages of each region. Map adapted from [Reddy, 2007].

Table B.3: F_{ST} estimates between the Dravidian Influenced and Non-Dravidian Influenced populations and the other continental populations.

Population	Non-Dravidian Influenced	Dravidian Influenced
CEU	0.0256	0.0496
JPT-CHB	0.0764	0.0806
YRI	0.1444	0.1474
East Asia	0.0772	0.0808
Europe	0.0227	0.0457
Mexico	0.0348	0.0492
Non-Dravidian Influenced	-	0.0121
Dravidian Influenced	0.0121	-

languages represented. For categorization purposes, we note that Malayalam and Tamil are Dravidian languages, andKonkani and Sinhalese both have borrowed words from Dravidian languages [Emeneau and Burrow, 1962], and we therefore group these languages into a "Dravidian Influenced" group. The six remaining languages we simply term as "Non-Dravidian Influenced" in subsequent analyses (Supplementary Table B.2). The Dravidian Influenced languages are predominately spoken in southern India (Supplementary Figure B.3).

We ran *STRUCTURE* using the same parameters as the global analysis but using 315 individuals from India and Sri Lanka having excluded individuals with no language information, have English as a primary language or are related. We used 6,542 SNPs having MAF > 0.2 (within South Asia) and separation of at least 400kb. Individuals were classified by self-reported language spoken. We excluded individuals without language information and those whose primary self-reported language was English. The results are shown in Supplementary Figure B.2C. At K = 2, there is no structure consistent with language

groups. However, at $K = 3$, we note that languages spoken in the south of India and Sri Lanka, including the Dravidian languages Malayalam and Tamil cluster together, as well as some Gujarati individuals. Sinhalese and Tamil are the two officially recognized languages of Sri Lanka. Malayalam is spoken along the tropical Malabar Coast of southwestern India, near Sri Lanka. Konkani is mostly spoken along the section of the south-western coastline of India known as Konkan, also near Sri Lanka. Further increasing the number of clusters to $K = 4$ increases admixture without any geographic or linguistic correlation.

Mexico: As discussed in the main text, we quantified the admixture in Mexicans using a *STRUCTURE* analysis of Mexicans, Europeans and East Asians. We extracted 778 individuals from the POPRES, comprising of 107 Mexican individuals, 400 randomly selected European individuals with known European grandparents and 271 East Asian individuals (including 90 HapMap individuals). We used 6,557 SNPs with MAF in these populations of > 0.2 (within Mexico) and spaced at least 400kb apart. The results are shown in Supplementary Figure B.2D. At $K = 2$, the Mexican individuals appear admixed between a predominately European cluster and a predominately East Asian cluster, with slightly greater membership in the former cluster. However, at $K = 3$, the Mexicans form their own cluster and no longer share East Asian admixture, but retain a 'European' admixture component. The average proportion of European admixture in Mexican individuals with $K = 3$ is 32.5% with a standard deviation of 17.4%. Further increasing K only reveals further admixture among European populations or separates the Japanese and Chinese populations.

We repeated the analysis using the 'supervised' *STRUCTURE* mode, having pre-assigned European and East Asian individuals to their respective popula-

tions. A $K = 3$, we found this method to give similar results to the unsupervised mode, with a European admixture component of 35.0% (standard deviation 16.8%) in Mexican individuals.

The first two principal components of the same individuals demonstrates a similar patter (Figure 2.1B), with Mexican individuals forming a distinct cluster between the European and East Asian Clusters in the first principal component. However, the second PC further differentiates the Mexican individuals from the East Asian individuals without substantially increasing the separation from Europeans.

Comparison with HGDP

While the global *STRUCTURE* analysis reveals broad patterns of population differentiation (Supplementary Figure B.2), the method is limited to using a small fraction of the available SNPs due to high computational cost. Furthermore, as the number of specified clusters is increased, the patterns of population structure become increasingly difficult to interpret. As an alternative means for analyzing population structure, we conducted a PCA of the genotype data [Patterson et al., 2006]. This method has the advantage of being able to analyze many more SNPs and can flexibly summarize patterns of both discrete [Patterson et al., 2006] and continuous spatial [Novembre and Stephens, 2008] population structure. PCA analysis of the POPRES alone is considered in Nelson et al [Nelson et al., 2008]. To investigate how the POPRES complements known patterns of global diversity, we combined the 2,943 "strict" individuals in the POPRES dataset with 479 individuals from the HGDP genotype

data [Jakobsson et al., 2008] for a combined total of 3,448 individuals. Although the two datasets were generated on separate genotyping platforms, more than 73,520 SNPs are shared even after pruning SNPs in high linkage disequilibrium (LD) and those with more than 5% missing data.

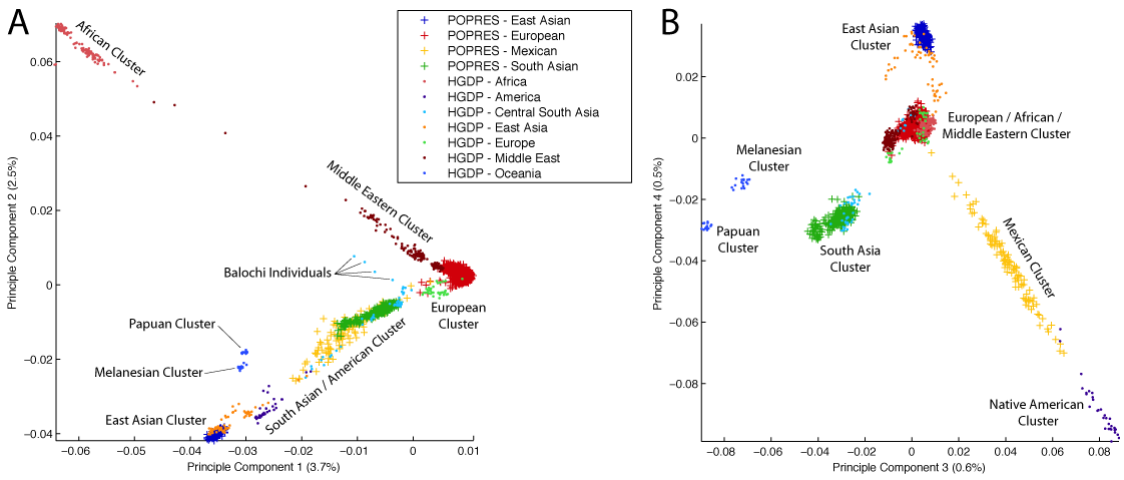

Figure B.4: (A) First two principal components of global PCA analysis using approximately 73,000 common SNPs from the POPRES and HGDP datasets. (B) Third and fourth principal components. The percentage of variance explained by each principal component is shown in brackets.

The first two principal components (PCs) of the combined dataset separate individuals into clusters largely determined by geographic origin (Supplementary Figure B.4A), which is consistent with a previous analysis of the HGDP dataset [Li et al., 2008]. Individuals from East Asia and Europe in the POPRES tend to cluster more tightly than those from the HGDP study. This is to be expected, as the POPRES samples are taken from presumably well-mixed urban populations whereas the HGDP sample is largely composed of diverse isolated populations (e.g. Basque, Sardinian, and Orcadians within Europe). Both the Mexican and South Asian individuals cluster between the European and East Asian clusters in this projection. The next two PCs reveal further structure within the Asian / American clusters, separating the Asian individuals from the American individuals (Supplementary Figure B.4B). Notably, the POPRES Mexican individuals form a new cluster between the predominately European cluster and the Native American cluster, which is indicative of the historical admixture of Europeans with Native Americans.

Phasing of the Data

For the estimation of haplotype diversity and population recombination rates, we first used the program *BEAGLE* version 2.1.3 to phase the genotype data [Browning and Browning, 2007]. This method was chosen as it is currently one of the few available methods that can phase adataset of this size in a reasonable time. Each sub-continental population in the strict dataset was phased separately. The default parameters were used with the exception of the European samples, for which we set nsamples =1 as recommended in the documentation for large samples. We phased the X chromosome separately, using an unpublished version of *BEAGLE* (version 2.2.0) that makes use of the known phase of the male samples.

Haplotype Diversity

To test whether the mean of the distribution of the number of haplotypes is informative of recent population demography, we conducted coalescent simulations using *ms* [Hudson, 2002]. We considered a family of demographic models (Supplementary Figure B.5) where in the present day there are two separate subpopulations, one of size N_c = 10,000 and the other of size N_c = 5,000. These two subpopulations do not exchange any migrants. Going back in time, at τ years ago, the two populations join and form an ancestral panmictic population of size N_a = 10,000. We examined a range of four different values of τ (0, 5,000, 10,000, and 20,000 years ago) for the population split times. To match our observed data, we sampled 146 chromosomes from eachsubpopulation and simulated 5,000 independent 500kb regions with an average per-generation re-

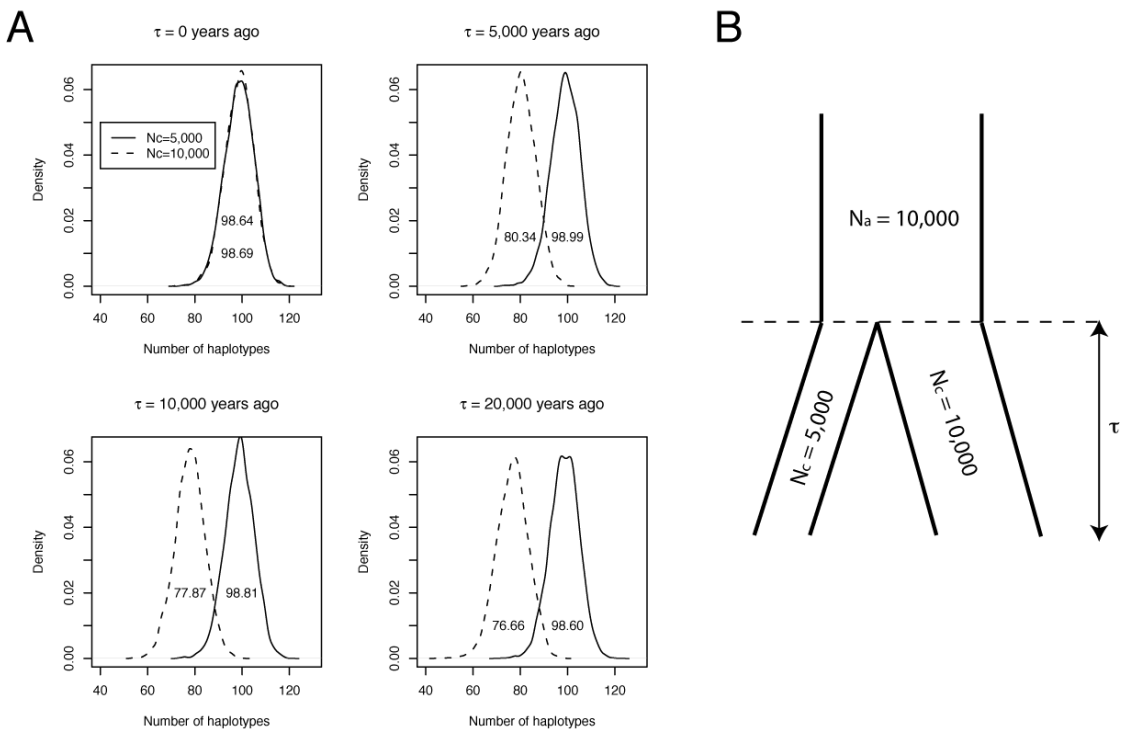

Figure B.5: (A) Distribution of the number of haplotypes for different population split times. The number inside each of the density plots is the mean of the distribution of the number of haplotypes. (B) Illustration of the population demography used in the simulations.

combination rate of 1cM/Mb. The *ms* command line used for these simulations is:

./ms 292 5000 -t 300 -r 200 500001 -I 2 146 146 0 -en 0 2 0.5 -ej τ 2 1 -F 29

where τ varies between simulations. Note that the mutation rate is set to be an arbitrary value, and does not matter given our sampling strategy of selecting a subset of SNPs (see below). We converted τ from generations to years assuming 20 years per generation.

In our analysis of the observed data, we only considered SNPs with MAF > 10% in all subpopulations. We implemented a similar filtering strategy in

Table B.4: Percentage of HapMap YRI haplotypes found in the European sample. This table is based on 25 SNP haplotypes in 2,925 windows of 0.5cM. The data was thinned to 114 chromosomes in each populations (to equal the YRI sample size).

Population	**Percentage of YRI haplotypes shared**	**Lower 95% C.I.**	**Upper 95% C.I.**
Europe (SW)	5.24%	5.03%	5.45%
Europe (S)	4.90%	4.70%	5.10%
Europe (C)	4.74%	4.53%	4.94%
Europe (NW)	4.73%	4.53%	4.94%
Europe (W)	4.72%	4.52%	4.93%
Europe (SE)	4.71%	4.51%	4.91%
Europe (NNE)	4.66%	4.46%	4.87%

Table B.5: Percentage of Mexican haplotypes shared with European populations. This table is based on 25 SNP haplotypes in 2,925 windows of 0.5cM. The data was thinned to 152 chromosomes in each populations (to equal that of the smallest European sample in the table, Europe NNE).

Population	**Percentage of Mexican haplotypes shared**	**Lower 95% C.I.**	**Upper 95% C.I.**
Europe (SW)	26.43%	26.05%	26.81%
Europe (S)	26.31%	25.92%	26.69%
Europe (W)	26.29%	25.92%	26.66%
Europe (NW)	26.14%	25.78%	26.51%
Europe (C)	26.12%	25.75%	26.49%
Europe (NNE)	25.93%	25.55%	26.30%
Europe (SE)	25.85%	25.48%	26.23%

our simulations. In each simulation replicate, we selected a subset of 25 SNPs withMAF > 10% in both of the subpopulations. We selected the same set of SNPs for each subpopulation. Using these SNPs, we parsed the haplotypes found in each subpopulation and then counted the number of haplotypes in each subpopulation for each of the 5,000 simulation replicates.

Supplementary Figure B.5 shows the results of this analysis. Note that if the two populations (going backwards in time) joined immediately ($\tau = 0$), we do not see a difference in the distribution of the number of haplotypes between the

Table B.6: Estimates of Haplotype Diversity using a thinned sample of 40 chromosomes per population. High values within each continent are shown in bold. Confidence intervals for the haplotype counts are calculated assuming a normal distribution.

Population	H_{10}	**95% Confidence Interval**	H_{25}	**95% Confidence Interval**
Non-Dravidian Influenced	**33.5341**	33.365, 33.703	**22.4679**	22.246, 22.69
Dravid Influenced	33.4043	33.233, 33.576	22.3368	22.117, 22.556
Europe (NW)	31.213	31.026, 31.4	20.0524	19.836, 20.268
Europe (C)	31.5891	31.406, 31.772	21.0207	20.806, 21.235
Europe (NNE)	31.5986	31.419, 31.778	21.0613	20.848, 21.274
Europe (W)	31.6785	31.494, 31.863	21.263	21.056, 21.47
Europe (SE)	32.0286	31.849, 32.208	21.3254	21.11, 21.54
Europe (SW)	32.2328	32.056, 32.41	21.5581	21.34, 21.776
Europe (S)	**32.5165**	32.337, 32.696	**21.5941**	21.375, 21.813
Mexico	**31.3765**	31.202, 31.551	**20.9565**	20.743, 21.17
Japan	30.6862	30.489, 30.884	19.6953	19.479, 19.912
Taiwan	**31.3138**	31.118, 31.51	**20.7644**	20.553, 20.976

two populations. However, for the other values of τ, we consistently see that for the smallersubpopulation (dotted lines), the distribution of the number of haplotypes is lower than that for the larger population (solid lines). We also see that as the time since the population split increases, the smallersubpopulation has fewer and fewer haplotypes (compare τ = 5,000 years to τ = 20,000 years) as expected. These results suggest that the distribution of the number of haplotypes can be informative about recent demographic history.

In the main text, we analyzed populations with at least 73 individuals. For this reason, the Dravidian Influenced group was not included. However, using a thinned sample of 20 individuals per population, we were able to compare

the Dravidian Influenced group to the other populations B.6. We see that the two South Asian populations have similar levels of haplotype diversity. For the other populations, the relative levels of diversity are nearly identical to the analysis using 73 individuals.

To understand how the haplotype diversity statistics are influenced by SNP ascertainment bias, we conducted additional coalescent simulations using the same two-population split model with τ fixed at 20,000 years. We simulated a genotype sample of 146 chromosomes from each population and a SNP discovery sample of four chromosomes in each population. The two genotype samples did not include any of the chromosomes used for SNP discovery. We considered four different ascertainment strategies relevant for the Affymetrix 500k data: 1) only considering SNPs polymorphic in two discovery chromosomes from the smaller population, 2) only considering SNPs polymorphic in four discovery chromosomes from the smaller population, 3) only considering SNPs polymorphic in four chromosomes from the larger population or the smaller population (e.g. using four SNP discovery chromosomes from each population), and 4) complete ascertainment in both populations. These ascertainment strategies are meant to mimic the actual ascertainment process where the genotyped SNPs are likely to be at high frequency due to discovery in a small number of chromosomes. Equally important, we considered differences in SNP discovery between populations, as SNP discovery was not uniform across all the populations considered in our study (e.g. little or no SNP discovery has been conducted in the South Asian population).

We simulated a single set of 5,000 independent regions and then implemented the four ascertainment strategies described above. Any differences in

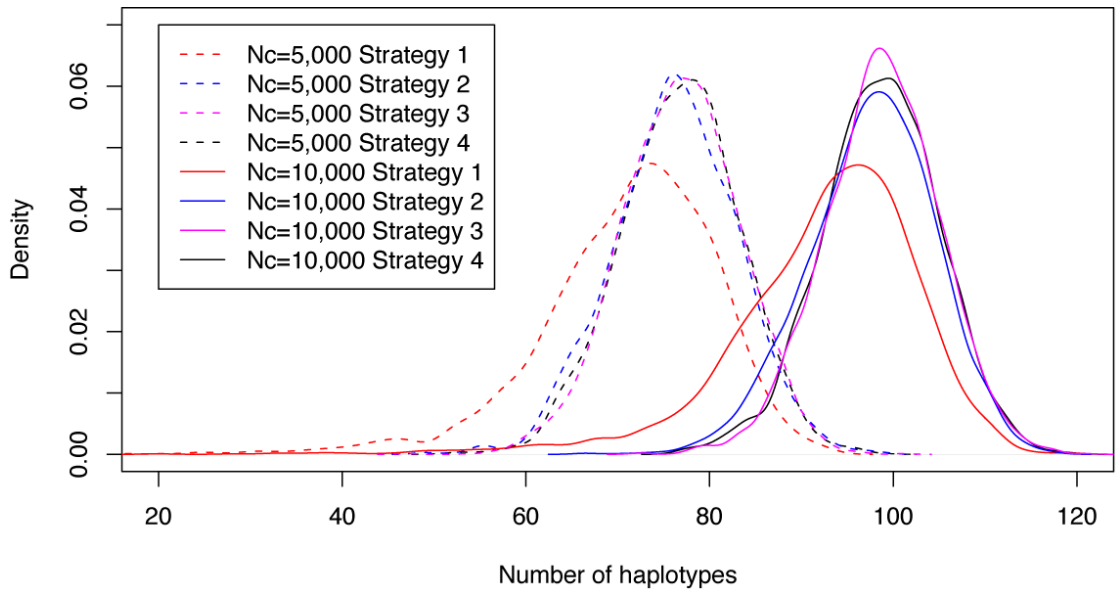

Figure B.6: The effect of SNP ascertainment on the distribution of the number of haplotypes. We considered four different ascertainment strategies: 1) SNPs polymorphic in two discovery chromosomes from the smaller population (red lines), 2) SNPs polymorphic in four discovery chromosomes from the smaller population (blue lines), 3) SNPs polymorphic in four chromosomes from the larger population or the smaller population (e.g. using four SNP discovery chromosomes from each population; pink lines), and 4) complete ascertainment in both populations (black lines). Dotted lines represent the distribution of the number of haplotypes for the smaller population (N_c = 5,000) and solid lines the distribution of the number of haplotypes for the larger population (N_c = 10,000).

the distribution of the number of haplotypes among ascertainment strategies are therefore not due to the evolutionary variance among different coalescent simulation replicates, as the same simulation replicates were used for all ascertainment strategies. For each region, we selected a random subset of 25 SNPs withMAF > 10% in both populations. As in our analysis of the real data, the same set of SNPs was used in both populations. Importantly, haplotypes under each ascertainment strategy all consist of 25 SNPs. Thus any differences in the number of haplotypes among different ascertainment strategies are not due to the fact that we are missing many SNPs when a small SNP discovery sample was used.

Supplementary Figure B.6 shows the distribution of the number of haplotypes for the small (N_c = 5,000; dotted lines) and in the large (N_c = 10,000; solid lines) populations for the four different ascertainment strategies. For all four ascertainment strategies, we see that the distribution of the number of haplotypes is higher for the larger population, indicating that haplotype diversity is related to population size, even when there is no SNP discovery from the larger population. While the overall means of the distributions appear quite similar regardless of ascertainment strategy, the distributions do differ for different ascertainment strategies. For example, using only two chromosomes from the smaller population for SNP discovery (ascertainment strategy 2, red lines in Supplementary Figure B.6) results in more regions with a smaller number of haplotypes for both populations. Increasing the number of SNP discovery chromosomes from 2 to 4 greatly reduces this problem (compare the blue lines to the red lines). These simulations, in agreement with previous empirical evidence [Conrad et al., 2006], suggest that qualitative patterns of haplotype diversity such as the number of haplotypes averaged over many windows of the genome are largely robust to ascertainment bias. We caution that other haplotype orLD statistics may be more sensitive to ascertainment bias and additional investigation of their properties may be warranted.

Identification of Runs of Homozygosity

To assess the robustness of the method to issues regarding SNP ascertainment, we conducted a simulation study using a similar scheme to that adopted for the haplotype diversity simulation study. Using the program *GENOME* [Liang et al., 2007], we simulated chromosomes of 5cM in two populations that

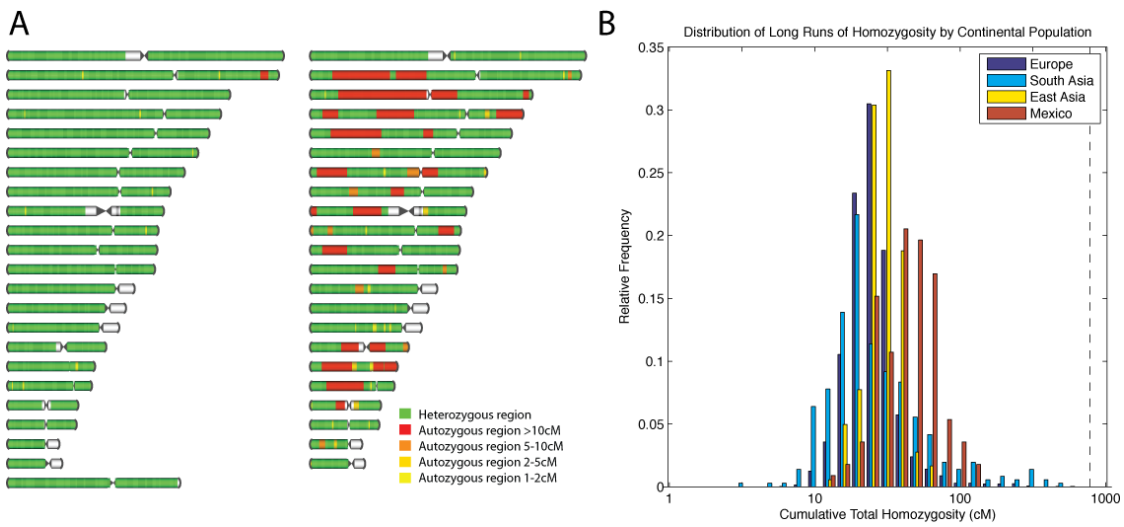

Figure B.7: HHRs in the four continental populations. Colors indicate the percentage of individuals with LROHs in each region of the genome.

separated 1,000 generations ago. As before, the ancestral population had an effective population size of 10,000, and the two sampled populations had effective population sizes of 10,000 and 5,000. We set the recombination rate to be equal to 1cM/Mb and the mutation rate to 1 x 10^{-8} per bp. We randomly combined pairs of simulated chromosomes to create simulated individuals. By chance, some of these individuals will have regions of autozygosity, and we tested the robustness of the method to detect these regions under a variety of SNP ascertainment schemes.

Using the unthinned simulated data (with approximately 7,000 to 8,000 SNPs per simulation), we estimated the cumulative LROH in each individual (cROH) without ascertainment of any kind. We selected 253 and 115 individuals from the small and large populations respectively with more than 1cM cROH. We then created 4 simulated data sets using different SNP discovery schemes: 1) SNPs discovered in a panel of 4 chromosomes from the large population, 2) SNPs discovered in a panel of 4 chromosomes from the small population, 3)

Table B.7: Robustness of HMM method to SNP ascertainment. The table shows the correlation between cROH estimated with full SNP discovery compared to the cROH estimated under 4 other ascertainment schemes. For comparison, a similar study was performed using F. Only simulated individuals with cROH > 1cM (estimated under full ascertainment) were used in the calculations. Both F and cROH were estimated using within-population SNP frequencies.

	Small Population (N=253)		Large Population (N=115)	
Ascertainment Scheme	cROH	F	cROH	F
Complete Ascertainment	1.0	1.0	1.0	1.0
4 chr from large population	**0.966**	0.815	**0.994**	0.911
4 chr from small population	**0.974**	0.845	**0.987**	0.921
4 chr from each population	**0.974**	0.921	**0.993**	0.952
2 chr from each population	**0.978**	0.872	**0.994**	0.928

SNPs discovered in 2 chromosomes from each population, 4) SNPs discovered in 4 chromosomes from each population. Once all the SNPs had been ascertained, we further thinned to 1,000 SNPs that approximately match the mean genetic distance between SNPs and frequency spectra observed in our study using the Affymetrix 500K chip. This was achieved by first constructing a site frequency spectrum of both our observed data and the simulated data. We then repeatedly removed SNPs from over-represented frequency classes until only 1,000 SNPs remained in the simulated data set.

We re-estimated cROH using the ascertained data. Robustness of the method to SNP ascertainment was measured by calculating the correlation coefficient between thecROH estimate using the unthinned data and the estimate under the ascertainment scheme for all individuals with more than 1cM of cROH . For

Table B.8: Regions appearing to be LROH in over 10% of individuals within a population.

Chr	Start	End	Region	# SNPs (MAF > 5%)	Mean % of Individuals with LROH	Population
1	15379784	17401898	1p36.13	136	13.5	East Asia
2	3045705	3710421	2p25.3	68	10.3	Mexico
2	8688612	10000083	2p25.1	116	22.1	East Asia
2	8899545	9854486	2p25.1	109	12.8	Mexico
2	43179074	44202272	2p21	94	11.3	East Asia
2	157991956	159379990	2q24.1	138	10.9	Mexico
2	176871680	177857610	2q31.1	94	13.9	East Asia
2	205483512	206252910	2q33.3	130	11.5	Mexico
3	43179088	45124712	3p21.33	161	10.1	East Asia
3	121522065	122818151	3q13.33	124	10.4	East Asia
3	189688953	190302800	3q28	60	11.6	East Asia
3	198067604	198958007	3q29	67	11.4	East Asia
4	29372445	29998717	4p15.1	54	11.5	Mexico
4	32250599	34658227	4p15.1	181	26.0	Europe
4	32250599	34826055	4p15.1	201	12.1	Mexico
4	32528188	34431234	4p15.1	166	12.7	South Asia
4	32555448	34431234	4p15.1	138	19.0	East Asia
4	40844073	41888547	4p13	120	10.2	Mexico
4	41017949	42342777	4p13	110	22.2	East Asia
4	158335214	160167630	4q32.1	153	11.4	East Asia
5	116125903	118646439	5q23.1	187	11.0	East Asia
6	105599748	106475582	6q21	99	10.6	Mexico
8	10509878	12039387	8p23.1	169	17.5	East Asia
8	10509878	12039387	8p23.1	260	11.5	Mexico
10	21519891	23314154	10p12.31	64	10.4	East Asia
13	18441915	19690082	13q12.11	116	10.1	Mexico
15	61189627	64122891	15q22.31	173	16.3	East Asia
16	17231173	17878102	16p12.3	68	12.0	East Asia
16	68106289	71557266	16q22.3	288	10.2	Mexico
17	53118270	54758734	17q22	91	15.8	East Asia
21	15813718	16718760	21q21.1	90	16.4	East Asia
22	34790020	35312621	22q12.3	58	10.3	East Asia
22	37049908	37855737	22q13.1	60	11.9	Mexico
22	44385321	45441994	22q13.31	51	11.4	East Asia
X	47266602	57222190	Xp11.22	222	13.7	Mexico
X	100386133	111121991	Xq22.3	342	12.6	Mexico
X	106862626	111770922	Xq22.3	123	32.1	East Asia
X	146603508	148146339	Xq28	65	17.0	East Asia
X	146603508	148146339	Xq28	94	14.4	Mexico

comparison, we estimated a similar correlation coefficient using the inbreeding coefficient of these individuals, F, as estimated by *PLINK*.

We find that the HMM is largely robust to ascertainment scheme. In absolute terms, the estimated cROH under each ascertainment scheme was within 2% of the value estimated using the unthinned data. The correlation between the ascertained estimate and the unthinned estimate is very high (Supplementary Table B.7), especially in comparison to F. Depending on the ascertainment

scheme, the cROH method has correlation coefficients in the range of 0.966-0.978 for the small population, and 0.985-0.994 for the large population. In comparison, the F method has correlation coefficients in the range of 0.815-0936 for the small population, and 0.911-0.954 for the large population.

A potential confounding factor in the detection of LROHs using SNP genotype data is that SNPs occurring within copy number variable regions may appear to be homozygotic. For example, a hemizygous deletion of a region containing a SNP would potentially cause the SNP to be called as a homozygote. For this reason, we have attempted to remove SNPs within hemizygous regions by analyzing samples for copy number variable regions. To locate regions of hemizygous deletion we used the *CNAT 4.0* copy number tool command line version (Affymetrix). Individual CEL files were normalized using the quantile normalization method. One hundred random females were used to generate the pooled reference sample and *CNAT 4.0* was run with the Gaussian smoothing option on and band width set to 100kb. All other options were set to default. Hemizygous deletion regions were then called for each individual as regions showing 3 or more SNPs in the hemizygous state with p-value less than 10^{-3}. Further, regions called hemizygous which contained large gaps in SNP coverage or lowsnp density were removed before comparison to the regions of autozygosity.

To assess if we would expect to observe Highly Homozygous Regions (HHRs) under a standard coalescent model, we simulated 100 datasets using the program *ms* [Hudson, 2002], each consisting of 20Mb regions in 250 individuals. Simulations were conducted with an effective population size of 10,000 with a 90% bottleneck between 1,600 and 2,400 generations in the past. We used a pop-

ulation mutation rate (θ) of 400 / Mb, and a recombination rate of 1cM/Mb (ρ = 400 / Mb). Using this simulated data, we applied the HMM method and called LROH over 1cM in length and containing at least 50 SNPs. We then looked for regions where the LROH of overlap in 5% or more individuals. We found 7 of the 100 simulations contained regions of LROH in more than 5% of individuals. However, these regions we all below 0.85Mb in length and no region was homozygous in more than 6% of individuals. We therefore suggest that very long HHRs occurring at high frequency are either explained by a stronger bottleneck that that simulated here, or are indicative of usual and localized ancestral histories.

We considered the possibility that HHRs contain large inversions. Recombination is expected to be repressed within inversions [Stefansson et al., 2005], and hence would not break down the linkage between SNPs on the inversion haplotype. Assuming a given inversion reaches intermediate frequency in a population, then a fraction of individuals are likely to be homozygous at the inversion locus. However, comparisons with published lists of inversions [Tuzun et al., 2005, Bansal et al., 2007] do not suggest that any of our top HHRs contain previously identified inversions.

APPENDIX C

SUPPLEMENTAL INFORMATION FOR CHAPTER 3

Data Quality Control

The HapMap II release 23 and PopRes samples were genotyped and called according to their respective quality control procedures [Frazer et al., 2007, Nelson et al., 2008]. For our data quality control, we took 225 African individuals from 11 populations and 8 European batch controls that were genotyped in Affymetrix 500K arrays using their standard procedure; genotypes were called by Affymetrix. Following Affymetrix quality control procedures any individuals with < 93% QC call rate on either the Nsp or Sty chips were excluded, as well as any individuals showing low concordance between replicated markers on both chips. Subsequent to genotype calling, 3 individuals were excluded for low call rate. This resulted in 148 African individuals which passed these quality checks.

We exercised caution in assessing individual and genotype quality in our African samples. To avoid removing samples whose quality is acceptable except on markers which perform poorly in Africans, we chose to first remove markers which had low quality in Africans prior to removing individuals or arrays with overall poor quality. To ensure the high quality of our markers, we first removed 124,193 SNPs where the genotypes were called missing in > 5% of the African individuals. After removing these markers with this stringent cutoff, no individuals had >10% missingness so no further exclusions were made. Upon merging the African data with the YRI, African American, and European datasets from PopRes [Nelson et al., 2008], an additional 16,393 SNPs were removed to ensure no greater than 10% genotype missingness per marker. Further analyses revealed that two pairs of African individuals were likely related with identity by state (IBS) of > 0.81. One individual from each pair was removed for

suspected relatedness. A list of all populations included in the study and their sample sizes is found in Table C.1.

Preliminary analyses suggested systematic differences existed between the HapMap YRI genotypes and the 500K data. To ensure highest quality of analyses, we subsequently removed markers showing large differences between observed and expected Hardy-Weinberg equilibrium within the PopRes samples and within the African samples. We also excluded markers which were found to have allele flips relative to the HapMap study. Lastly, we removed markers with a minor allele frequency (MAF) < 0.01. This resulted in the removal of 12,980 additional SNPs resulting in a final dataset for all analyses in this study of 351,753 markers and 968 individuals. A complete summary of the QC procedure and exclusion of SNPs and individuals is in Figure C.1. In Figure C.2, we provide Frappe results for population structure analysis assuming $K = 2$ to $K = 7$ ancestral populations.

LD decay and comparison among African populations with varying sample size

Pairwise r^2 calculations were performed using the `--ld` command from *PLINK* [Purcell et al., 2007], with a minor allele frequency threshhold of 0.1, then were averaged across 500 bp bins. To adjust for variable sample sizes among our populations, we calculated the LD for each bin by randomly sampling 5 individuals from each population 100 times and averaging each bin across all 100 individual samplings. Wright's inbreeding coefficient, F, was calculated using the `--het` command in *PLINK* as the equation shown in [Purcell et al., 2007].

In order to properly compare the LD decay curves, LD was calculated for all populations at the minimum sample size, n, for any given population surveyed which was $2n = 10$ chromosomes for the Xhosa population. For all populations except Xhosa, the curve in Figure S3A is the mean of 100 random subsamples of size $2n = 10$ chromosomes. To assess the effect of not resampling on the pattern of LD decay comparing the Xhosa and the other populations, we also plotted LD for a single subsample of 5 individuals representing each other population and LD for the full sample of 5 Xhosa individuals (Figure C.3B) but overall the results remained unchanged, suggesting that resampling would not strongly influence the conclusions about the Xhosa decay curve.

Population structure inference

Since the International Haplotype Map project contained individuals of Yoruban ancestry from Ibadan, Nigeria, this population (YRI) has often served in medical and population genetics analyzes as a proxy for the ancestral population of African Americans. It is of critical interest to assess whether this assumption is valid. To assess the effect of using only one African population (the YRI in this case) in estimating African American ancestry, we compared results from *FRAPPE* analyses using all 12 African populations to the analysis using only the YRI (Figure C.4). Reassuringly, we found quite similar results for both K = 2 and K = 4 in estimates of African ancestry for African-Americans when the YRI are used alone or in combination with 11 additional populations ($r^2 \geq 0.9999$). It is important to note, however, that at K = 4 the clustering showed that while African Americans share large amounts of ancestry with the YRI, many (though not all) African American individuals also revealed a size-

able component of their ancestry from a source distinct from both European and YRI ancestry (shown in yellow, Figure C.4B). Furthermore, in agreement with previous haplotype analysis studies, the ancestry estimates of southern Europeans showed small amounts of shared ancestry with the African populations at this and larger values of K [Auton et al., 2009].

Several self-identified African Americans appear to have no recent African ancestry and others show no recent European ancestry. For example, Figure 3.3F summarizes the ancestry plot of an individual self-identified as African Americans who we have estimated has 99.8% European ancestry. Overall, we find approximately 1.3% of the African American individuals in our sample are estimated to have less than 1% African ancestry and 1.9% are estimated to have over 99% African ancestry as estimated by both *FRAPPE* and our PCA-based method.

We also used *FRAPPE* to study regional differences in admixture proportions among the African Americans sample. Grouping the African Americans by region, the lowest median African ancestries were in the Southwest (77.3%), Atlantic (78.8%) and West (79.3%), and slightly higher in the Midwest (80.5%). The highest median African ancestry was from African Americans in the South (83.4%). However, these differences are not statistically significant (Kruskal-Wallis test, p-value = 0.43). We also calculated Wright's inbreeding coefficient, F, as a measure of heterozygosity for the West African populations and the African Americans (Figure C.5). All the African populations show similar levels of diversity, with lower F values in the African Americans and little differentiation between the distributions of F values among regions. The lack of clear substructure of the African Americans and overall homogeneity of African ancestry

among regions of the US is likely the result of no assortative mating by ancestral geography among African Americans after they were brought as slaves to the Americas.

PCA based admixture estimation algorithm

The power of principal component analysis to distinguish major continental populations [Xu and Jin, 2008, Li et al., 2008, Tian et al., 2008, Nelson et al., 2008, Auton et al., 2009] has inspired us to develop a fast and efficient approach for generating marker-by-marker estimates of ancestry based on distances in PCA space. Our algorithm is similar in spirit to that proposed by Paschou and colleagues [Paschou et al., 2007] but has the added advantage of estimating both genome-wide (i.e., "average") ancestry as well as ancestry at each SNP along the genome ("local" ancestry proportions).

The closer a given individual is to the African centroid of the European-Africa axis of variation, the higher their African ancestry. Comparing estimates of ancestry from our PCA based approach to those generated by the Bayesian clustering algorithm STRUCTURE on a random subset of 5,000 SNPs in the POPRES data, we find a strikingly strong correlation (>99%) between our estimates and those of STRUCTURE using K = 2 ancestral populations for African Americanss in our data. A similarly high correlation (99.98%) is observed between our ancestry estimates and those of *FRAPPE*. A nice computational advantage of our method is that it runs nearly instantaneously in comparison to STRUCTURE, which can take weeks to run on even small subsets of the data.

Here, we describe details of our algorithm for estimating "local" or regional

genomic ancestry (i.e., the number of European or African chromosomes at a given location in the genome). Informally, we use a sliding window approach to average principal component loadings in a given genomic neighborhood for a given individual and compare this value to average loadings for individuals from the reference source populations. Formally, we first run principal component analysis on the admixed individuals (e.g., African Americans) and their potential ancestral populations (e.g., the Europeans from PopRes and the diverse West Africans presented in this study). For each individual i across each w-length SNP window k, calculate the local PCA score as:

$$\texttt{score}_{ik} = M'_{ik} \times e_k$$

where M'_{ik} are the normalized and scaled genotypes of the markers (i.e., "0","1", or "2" depending on allelic coding) in window k for individual i, e_k is the vector of loadings corresponding to the markers in window k. (In our notation w represents a user-defined constant that is the window length; in practice we have found that w = 10-20 SNPs provides an optimal trade-off between local information and smoothing for human data). The resulting data can further be modeled as a continuous-valued discrete stochastic process indexed by genomic location. For example, we have developed a three-state Hidden Markov model with hidden states (0,1,2) corresponding to the number of ancestral African chromosomes and the PCA scores as the "emitted" or observed signal. This allows us to further refine the PCA signal and provide a powerful means for local admixture estimation.

To estimate the number of European versus African chromosomes at each window, we use the Viterbi algorithm to find the most likely path between states, which represents the local individual ancestry estimates [Durbin et al., 1999]. States 0, 1, and 2 corresponding to the number of European

alleles are modeled as two independent two-state Markov chains, where each chromosome is an independent Markov chain either in state 0 or 1 European alleles. The transition probabilities, where $P(i, j)$ is the probability of transitioning from state i into state j, for each of the two-state markov chains are:

$$\begin{aligned} P(0,0) &= 1 - \pi(1-p) \\ P(0,1) &= \pi(1-p) \\ P(1,0) &= \pi p \\ P(1,1) &= 1 - \pi p \end{aligned}$$

where π is the probability of transitioning, a function of the inter-window distance, and p is the prior probability of an African allele, which we assume to be 0.8. We let

$$\begin{aligned} \pi &= 1.0 - e^{-2.0 * \gamma} \\ \gamma &= 0.001 * numGen * d \end{aligned}$$

where d is the distance between adjacent windows in centiMorgans as obtained from HapMap II, and *numGen* is the number of generations since admixture. We assume $numGen = 4.0$ as a conservative value, but have found our results to be robust to deviations from this value.

As seen in Figure 3.3B, the PCA-sliding window approach provides an intuitive and powerful visualization tool for admixture analysis. Specifically, note that the blue line corresponds to the mean PCA score for the African individuals in the data set along chromosome 1. Likewise, the red line is the mean PCA score for the European individuals in the POPRES data. The jagged black line is a single individual projected onto PC1 space. We see that the individual "cycles" between having 0,1, or 2 African ancestors at each genomic location

and this is evidenced by "tracking" either the European (red) line, a space between the red and blue lines (admixture), or the African (blue) line. The HMM can be thought of as a filter on the PCA genome-series data, and aids in local estimation (note that state "2" corresponds to tracking the African mean PCA, state "1" to the average of the African and European mean PCA, and state "0" to tracking the European mean PCA).

Validation of PCA based ancestry and application to demographic inference

In order to investigate the performance of our PCA-based ancestry algorithm, we created simulated admixed genotype data comparable to that analyzed in this study. To create each simulated admixed individual, we drew 128 African and European chromosome 22 phased haplotypes with approximately the same marker set analyzed in this study (4,279 SNPs) from the parental phased HapMap haplotypes. We drew haplotypes with 77% African and 23% European probabilities from the unrelated HapMap YRI and CEU individuals to make up the first generation of ancestors. We recombined the haplotypes once per generation, uniformly across the panel of markers, to simulate the next generation of admixed individual haplotypes, and repeated the admixture for 6 generations until we had 2 remaining haplotypes, which we then combined together into the resultant simulated individual's genotypes. We repeated this process to create 100 admixed individuals of approximately similar ancestry as our African American individuals, which we then ran with the YRI and CEU unrelated genotypes to estimate ancestry. We evaluated the performance of our

PCA-based admixture method on this small set of markers on chromsome 22. Across all individuals, our method had 96.3% accuracy per SNP per chromsome as compared to the true ancestry. To illustrate the accuracy of the method, we show several of the simulated individual's true ancestry on this chromosome as well as the ancestry estimated by our method in Figure C.6.

We have also validated our method empirically by using an F_{ST} based approach in which we compare the estimated degree of population differentiation among the populations used in the ancestry analysis of African Americans. Overall, we find that F_{ST} from autosomal markers between African and European populations is 13.9%. Likewise, autosomal F_{ST} between African Americans and Europeans was 9.6% and only 0.7% between African Americans and Africans. To test our approach, we perform a restricted analysis where we compare regions of the genome of each African American individual where both chromosomes were estimated to belong to one of the ancestral populations. Considering only the regions estimated to be of European ancestry within the African Americans, the F_{ST} between the Europeans and the "European segments of the genomes from African Americans" dropped to 0.06%. Likewise, using only the African-inferred regions of the genome for African Americans in our sample, the F_{ST} between these regions and the African populations decreased to 0.15%. We also calculated F_{ST} between the African only regions of the African Americans and each of the African populations. We found that the African populations with the smallest F_{ST} values with the African American regions were the non-Bantu Niger-Kordofanian speaking Igbo, Brong and YRI, each with with F_{ST} 's of less than 0.1% (Table C.2).

Our local individual ancestry estimates provide information useful for admixture mapping of African Americans. Since PCA results are affected by a number of factors including sample sizes, marker ascertainment, linkage disequilibrium between markers, and uniformity of markers chosen, our admixture estimates are also likely subject to these effects. Furthermore, PCA results are often difficult to interpret, and care should be used in applying our method to ensure that the PC axis corresponds to a clearly interpretable admixture component. Specifically, our method assumes that individuals appearing at intermediate values between two populations are admixed individuals, not simply individuals from a third distinct population. If these assumptions are upheld, we find that the PCA-based local admixture method gives consistent results in a very short amount of time; hence we expect that it will be applicable to large admixture studies where current approaches are time limiting. Furthermore, we expect that the PCA method should generalize well to multiple continental populations and provide accurate results when estimating admixture of individuals with ancestry from three or more distinct ancestral populations. It is also important to note that we attempted to use our PCA admixture method to infer within-continental ancestry assignment. However, the low degree of population differentiation within the African and European populations made it difficult to distinguish within continent ancestry in admixed individuals using a PCA of the genotypes.

Deviations from overall mean ancestry

Using our PCA-based method of determining ancestry, we found that ancestry at each location of the genome, averaged over all the African Americans in the

dataset, did not significantly differ from the genome-wide mean, with a range from 0.72 to 0.82 (Figure C.7). However, many regions showed moderate elevation or drops in mean African ancestry, including low (chr5p15, chr11q13) and high (chr6q12) mean African ancestry across our panel of African Americans. A complete list of regions, genes, and some of their associations is in Table C.4. Some interesting examples of genes within these regions of high or low ancestry include: ADAMTS-16, a protease expressed in ovarian follicles; BAI3, a brain-specific angiogenesis inhibitor that has been suggested by previous analyses to exhibit high allelic frequency differences between Africans and Asians [Hughes et al., 2008]; and TPCN2, a gene associated with pigmentation in Europeans [Sulem et al., 2008]. It is important to note that none of these regions are significant after correcting genome-wide for multiple testing. Nonetheless, our analyses raise the possibility that fast and inexpensive genotyping coupled with high-density ancestry reconstruction could be useful in identifying candidate genes under selection via an admixture selection mapping approach.

Web Resources

The PopRes individual genotypes and demographic data are available via the dbGaP archive sponsored by the National Center for Biotechnology Information. The HapMap individual genotypes are in release 23 of the International HapMap Project (`http://hapmap.org/`). The African data is available upon request.

Table C.1: Populations and sample sizes in study

Region	*n*	*Subpopulation (location)*	Latitude	Longitude	Language Group
hline Europe	400				
African American	365				
Africa	203				
	15	Bulala (Chad)	13.0	18.0	Nilo-Saharan (Sudanic)
	16	Kaba (Chad)	8.0	16.8	Nilo-Saharan (Sudanic)
	12	Mada (Cameroon)	10.8	14.1	Afro-Asiatic (Chadic)
	13	Hausa (Cameroon)	9.1	7.5	Afro-Asiatic (Chadic)
	13	Mbororo Fulani (Nigeria)	9.0	7.6	Niger-Kordofanian (non-Bantu)
	8	Brong (Ghana)	7.5	-2.0	Niger-Kordofanian (non-Bantu)
	17	Igbo (Nigeria)	6.0	7.0	Niger-Kordofanian (non-Bantu)
	57	Yoruba (Ibadan, Nigeria)	8.0	5.0	Niger-Kordofanian (non-Bantu)
	20	Bamoun (Cameroon)	5.5	10.8	Niger-Kordofanian (Bantu)
	18	Fang (Cameroon)	2.5	13.0	Niger-Kordofanian (Bantu)
	9	Kongo (D. R. C.)	-5.5	15.0	Niger-Kordofanian (Bantu)
	5	Xhosa (South Africa)	-32.0	28.0	Niger-Kordofanian (Bantu)

Table C.2: F_{ST} distances between major groups

Populations compared	F_{ST}	F_{ST}, African-only[1]	F_{ST}, European-only[2]
Africa - African Americans	0.7%	0.13%	14.0%
Europe - African Americans	9.6%	13.0%	0.06%
Africa - Europe	13.9%	-	-
Africa - Europe - African American	8.0%	-	-
Among African subpopulations	1.2 %	-	-

[1]Using the African-assigned regions of the African Americans only in the analysis
[2]Using the European-assigned regions of the African Americans only in the analysis

Table C.3: F_{ST} distances between African-only regions of the African Americans and each of the African populations, listed in ascending F_{ST} order

Population	Language Group	F_{ST}
Igbo	Non-Bantu Niger-Kordofanian	0.074 %
Brong	Non-Bantu Niger-Kordofanian	0.077 %
Yoruba	Non-Bantu Niger-Kordofanian	0.089 %
Kongo	Bantu Niger-Kordofanian	0.112 %
Bamoun	Bantu Niger-Kordofanian	0.201 %
Xhosa	Bantu Niger-Kordofanian	0.257 %
Fang	Bantu Niger-Kordofanian	0.266 %
Hausa	Afro-Asiatic	0.325 %
Kaba	Nilo-Saharan	0.353 %
Mada	Afro-Asiatic	0.970 %
Bulala	Nilo-Saharan	1.581 %
Fulani	Non-Bantu Niger-Kordofanian	2.973 %

Table C.4: Regions of high or low African ancestry and genes within the regions.

Region	Excess ancestry	Gene	Official Name
chr5: 3436080-6453181	European	IRX1	iroquois homeobox 1
5p15.3		ADAMTS16	ADAM metallopeptidase with thrombospondin type 1 motif, 16
		MED10	mediator complex subunit 10
chr6: 68710365-70587302	African	LMBRD1	LMBR1 domain containing 1
6q12		BAI3	brain-specific angiogenesis inhibitor 3
chr11: 67834242-68807621	European	MRGPRF	MAS-related GPR, member F
11q13		MRPL21	mitochondrial ribosomal protein L21
		CPT1A	carnitine palmitoyltransferase 1A (liver)
		MTL5	metallothionein-like 5, testis-specific (tesmin)
		LRP5	low density lipoprotein receptor-related protein 5
		TPCN2	two pore segment channel 2
		MRGPRD	MAS-related GPR, member D
		IGHMBP2	immunoglobulin mu binding protein 2
		GAL	galanin prepropeptide
		SAPS3	SAPS domain family, member 3

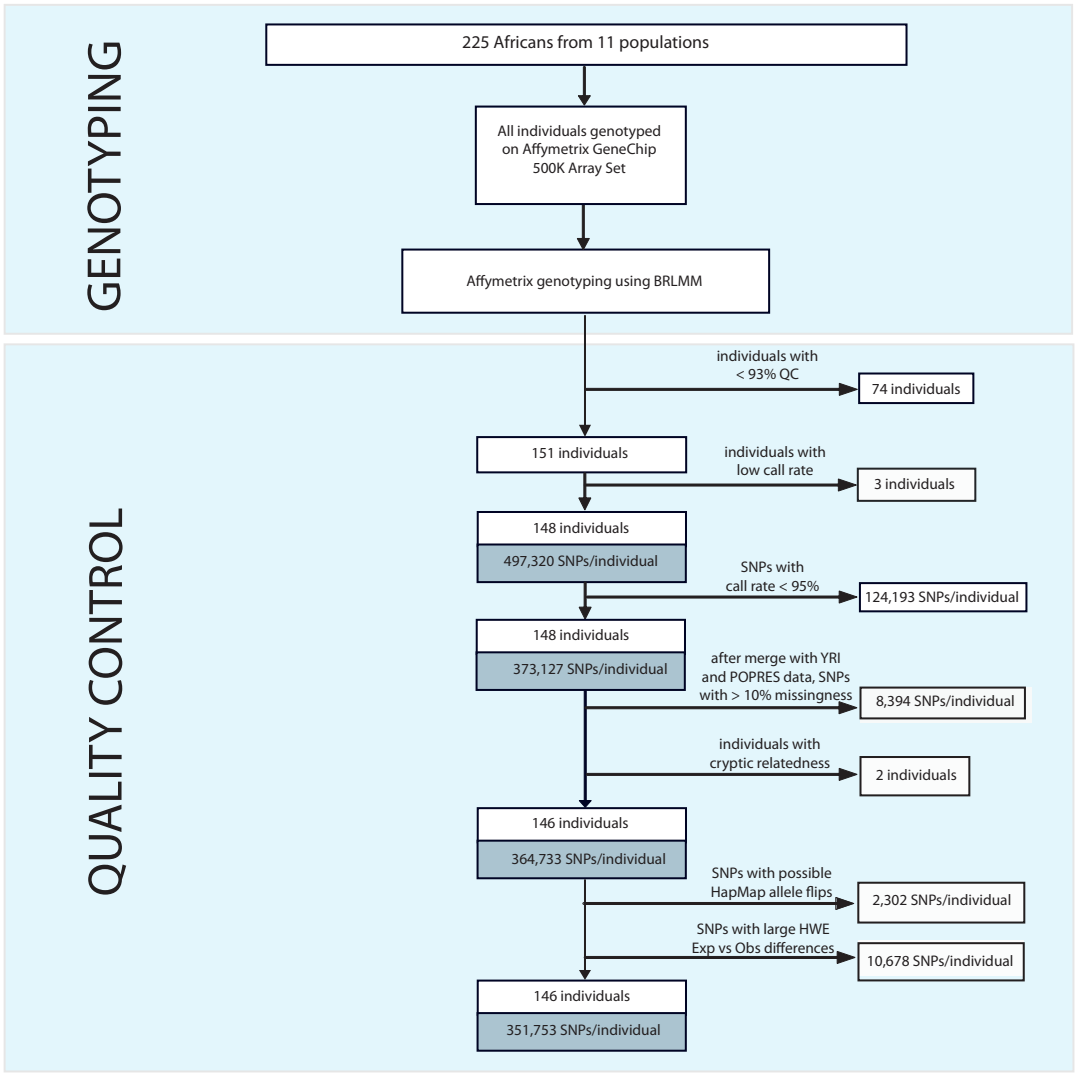

Figure C.1: Data genotyping and quality control process flowchart of inclusions and exclusions.

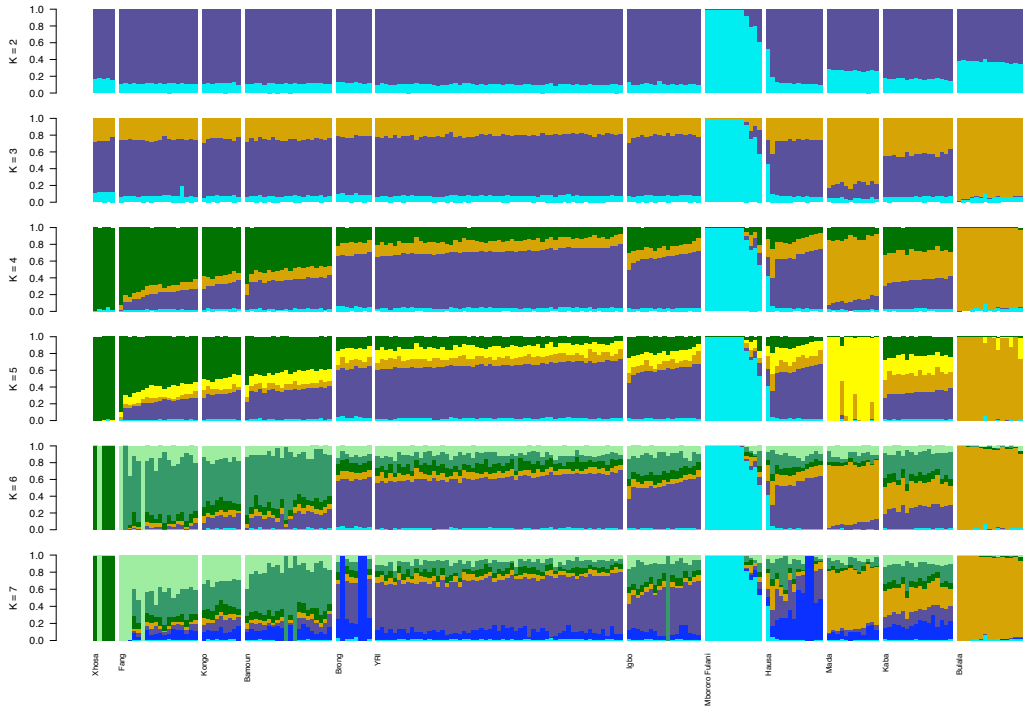

Figure C.2: *FRAPPE* analysis of the African populations. Individuals are represented as thin vertical lines partitioned into K segments corresponding to the inferred membership of the genetic clusters indicated by the colors. K, the prior number of clusters varies from $K = 2$ to $K = 7$

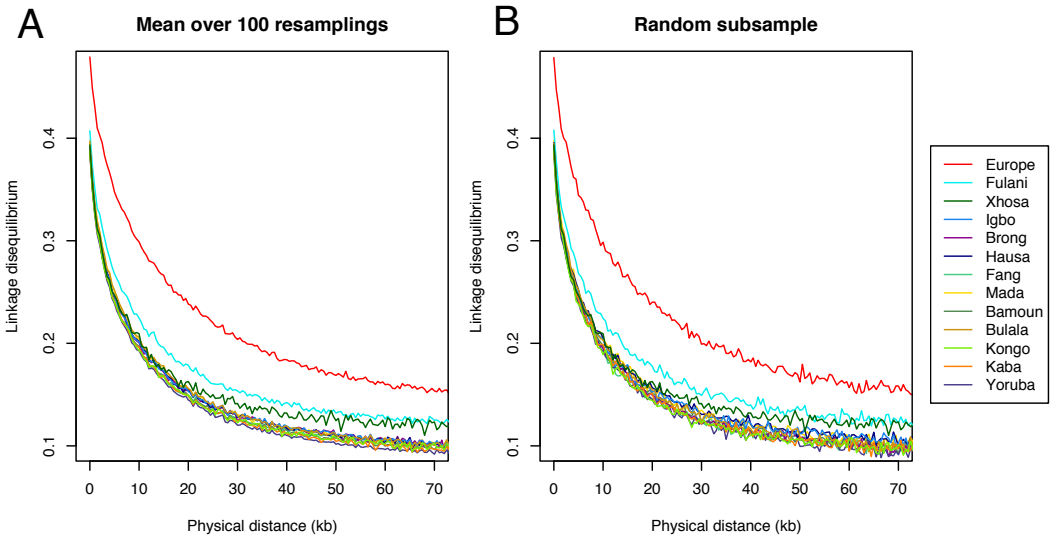

Figure C.3: Decay of linkage disequillibrium (in terms of r^2) for the populations studied. A. LD averaged over 100 iterations of sampling 5 individuals from each population. B. LD using a single random sample of 5 individuals from each population.

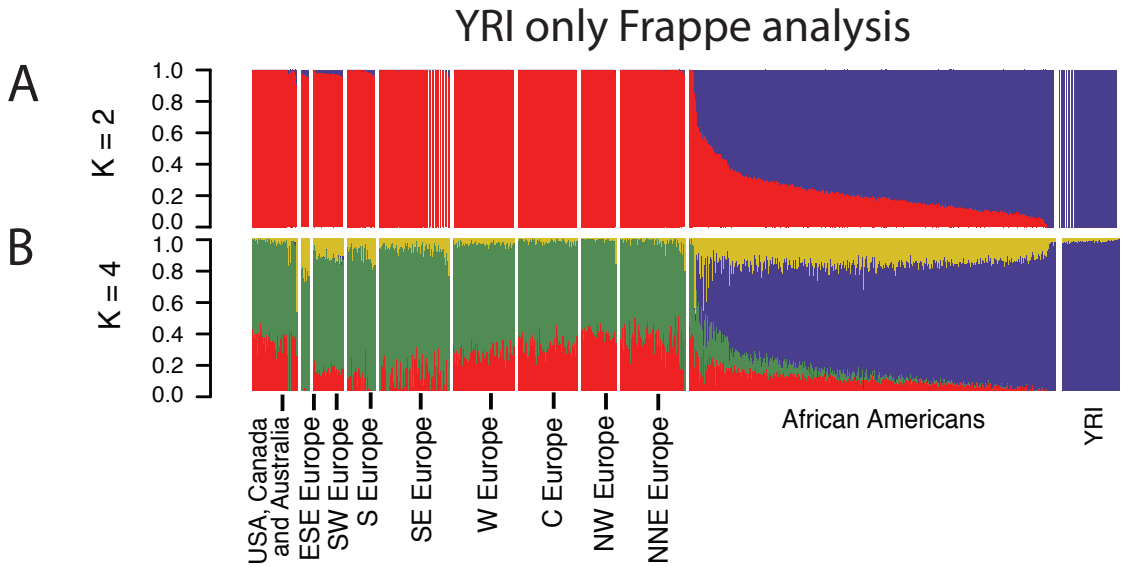

Figure C.4: *FRAPPE* clustering of Europeans, African Americans, and Yorubans. Individuals are represented as thin vertical lines partitioned into K segments corresponding to the inferred membership of the genetic clusters indicated by the colors. A) *FRAPPE* analysis with $K = 2$. B) *FRAPPE* analysis with $K = 4$.

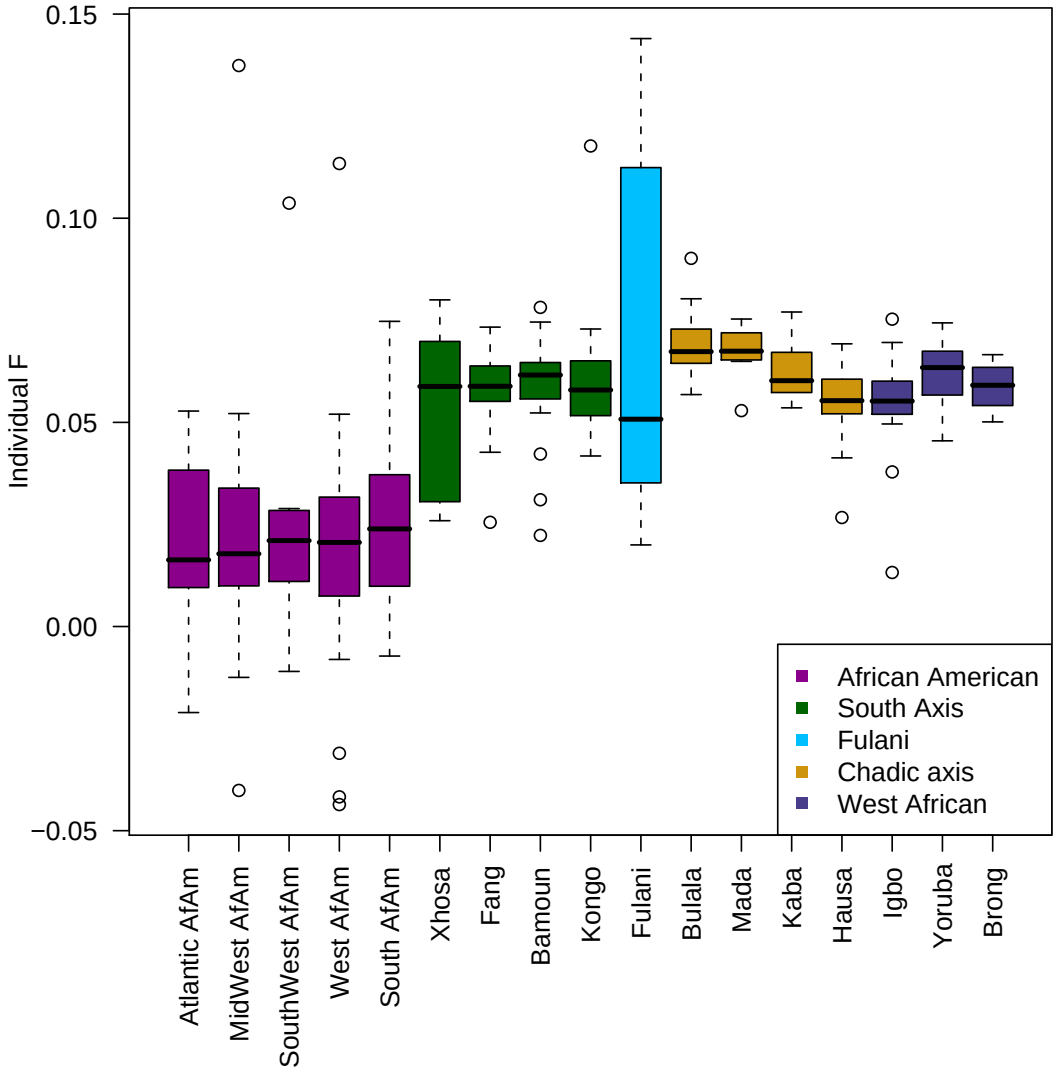

Figure C.5: Wright's inbreeding coefficient for individuals, grouped by population. African Americans (violet) are shown grouped by US region. The Atlantic region includes the Mid-Atlantic and New England states.Colors are derived from the population relatedness by clustering, and correspond to linguistic and geographic regions.

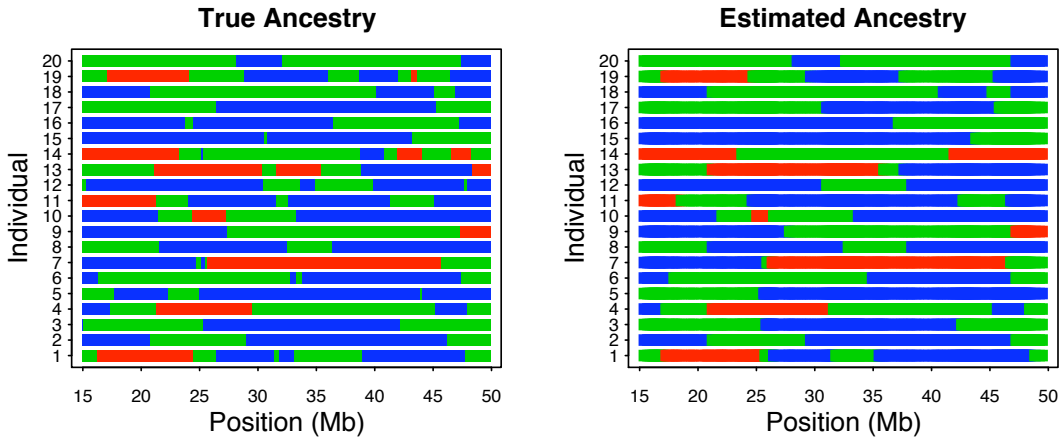

Figure C.6: True versus estimated ancestry for several simulated individuals along chromosome 22. True ancestry for a random sample of 20 of the simulated individuals (left). Ancestry estimates for these same individuals from the PCA-based local admixture method (right).

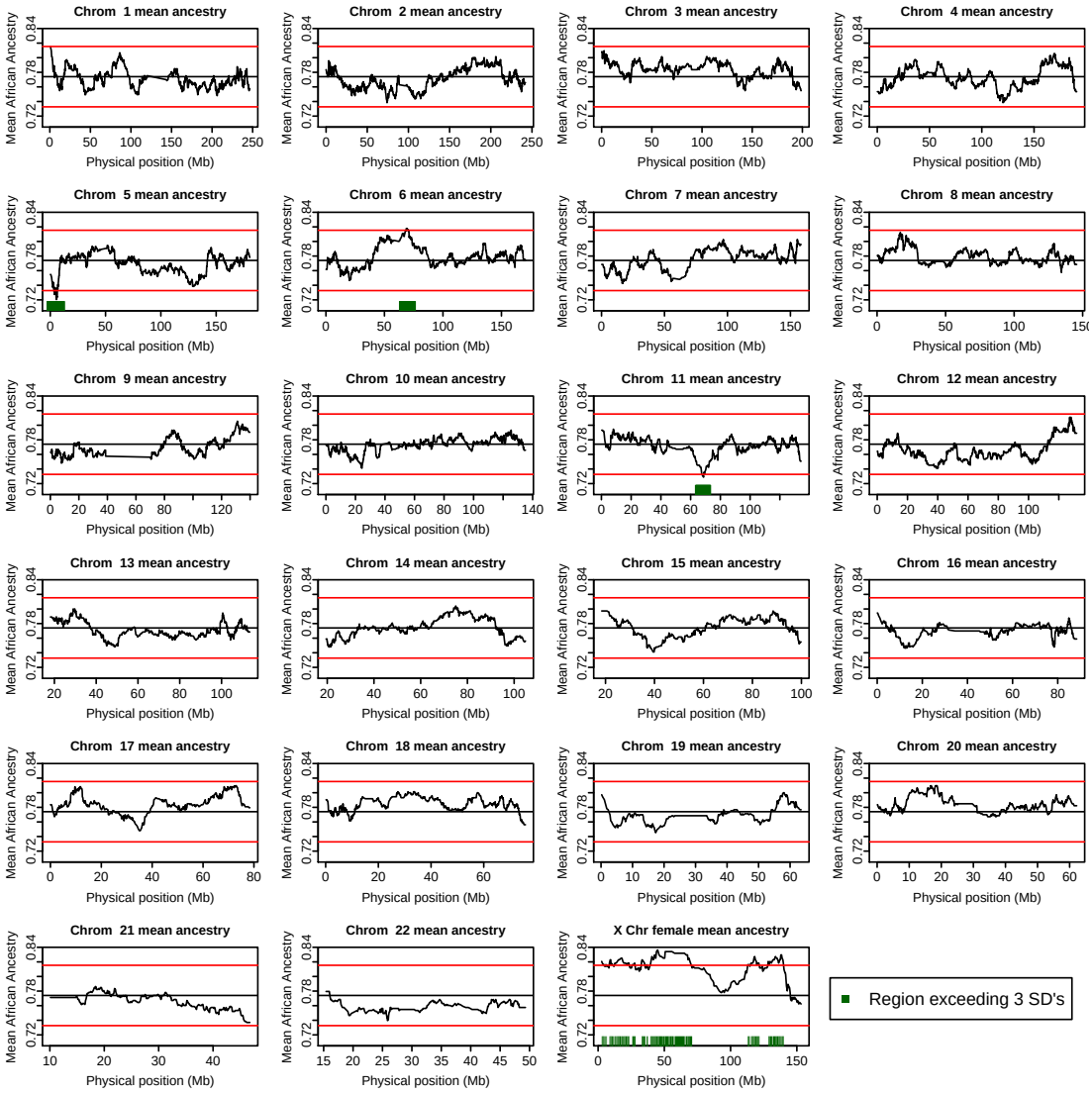

Figure C.7: Mean ancestry of 365 African American individuals at each window across each of the chromosomes. The black line shows the overall mean estimated ancestry. Red bands indicate + 3 and - 3 standard deviations from the mean ancestry.

APPENDIX D

SUPPLEMENTAL INFORMATION FOR CHAPTER 4

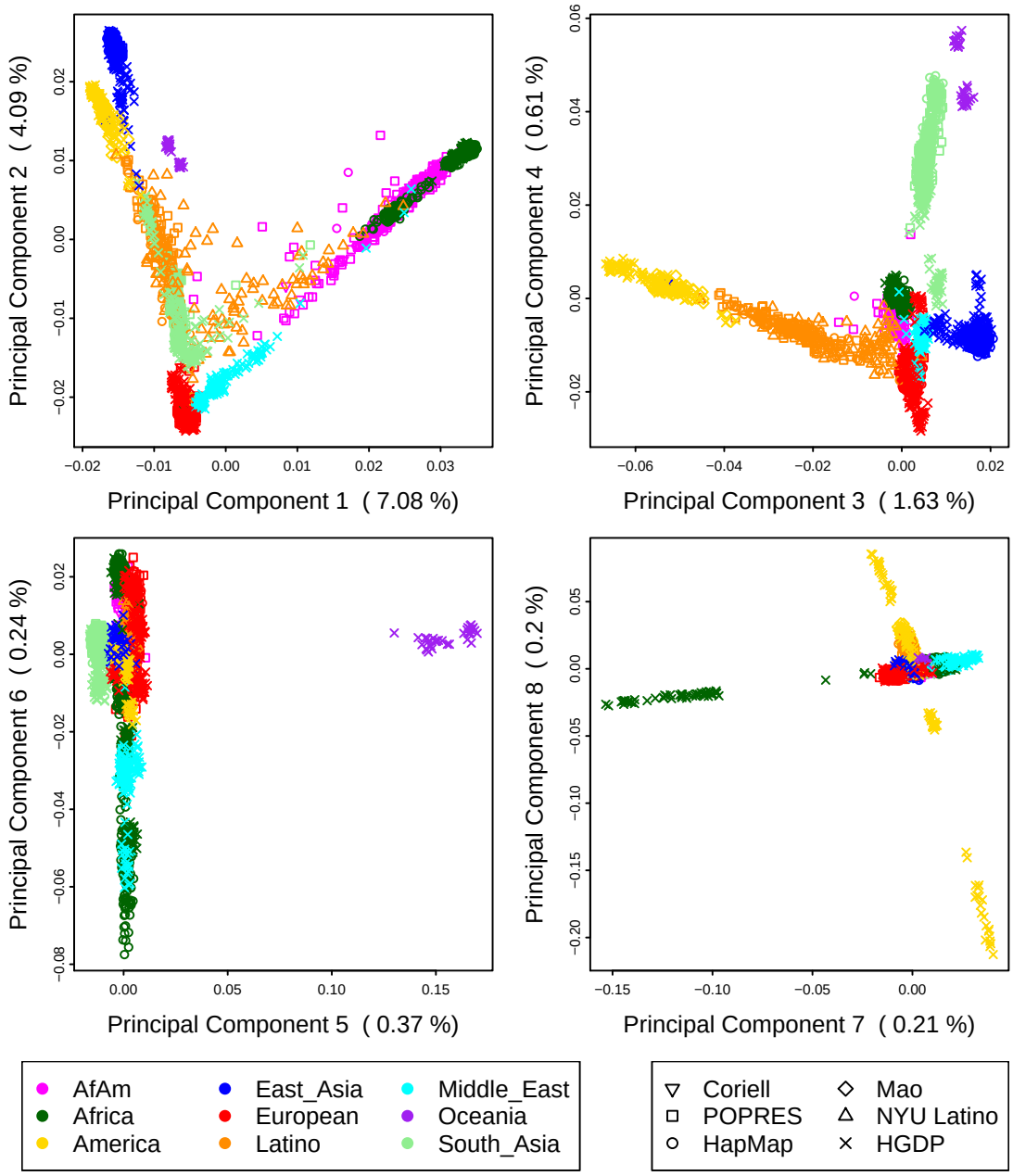

Figure D.1: Principal component 1 through 8 of all the individuals in the merged dataset, colored by population.

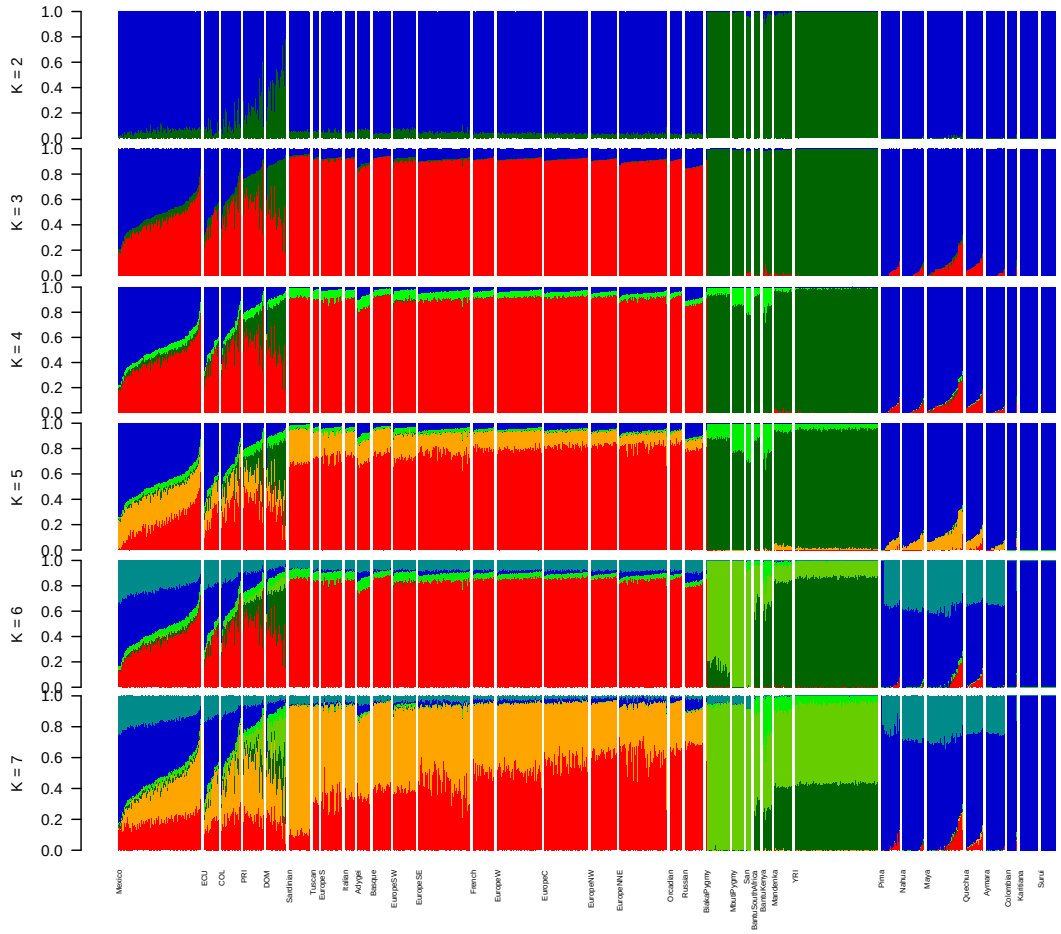

Figure D.2: Frappe clustering of the HL individuals as well as Europeans, Africans, and Native Americans, shown for $K = 2$ through $K = 7$.

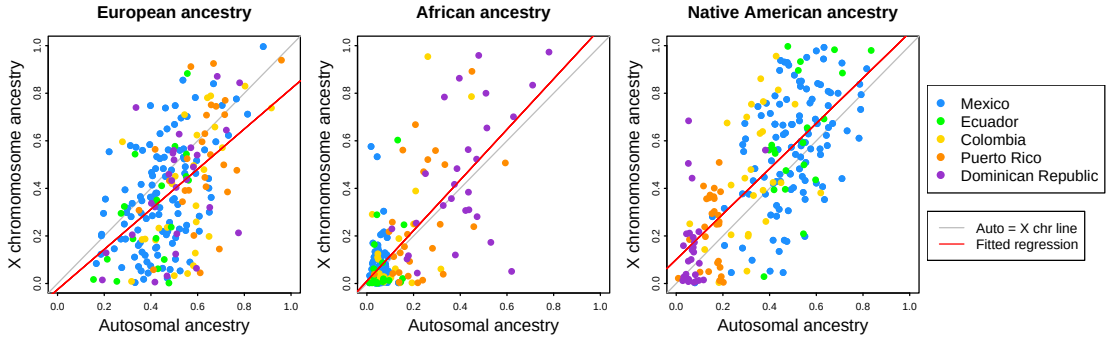

Figure D.3: Individual scatterplots comparing autosomal versus X chromosome ancestry proportions as estimated by *STRUCTURE* $K = 3$ using prior population information of ancestral populations. Expected autosome = X chromosome ancestry lines are show in gray, fitted linear regression lines in red.

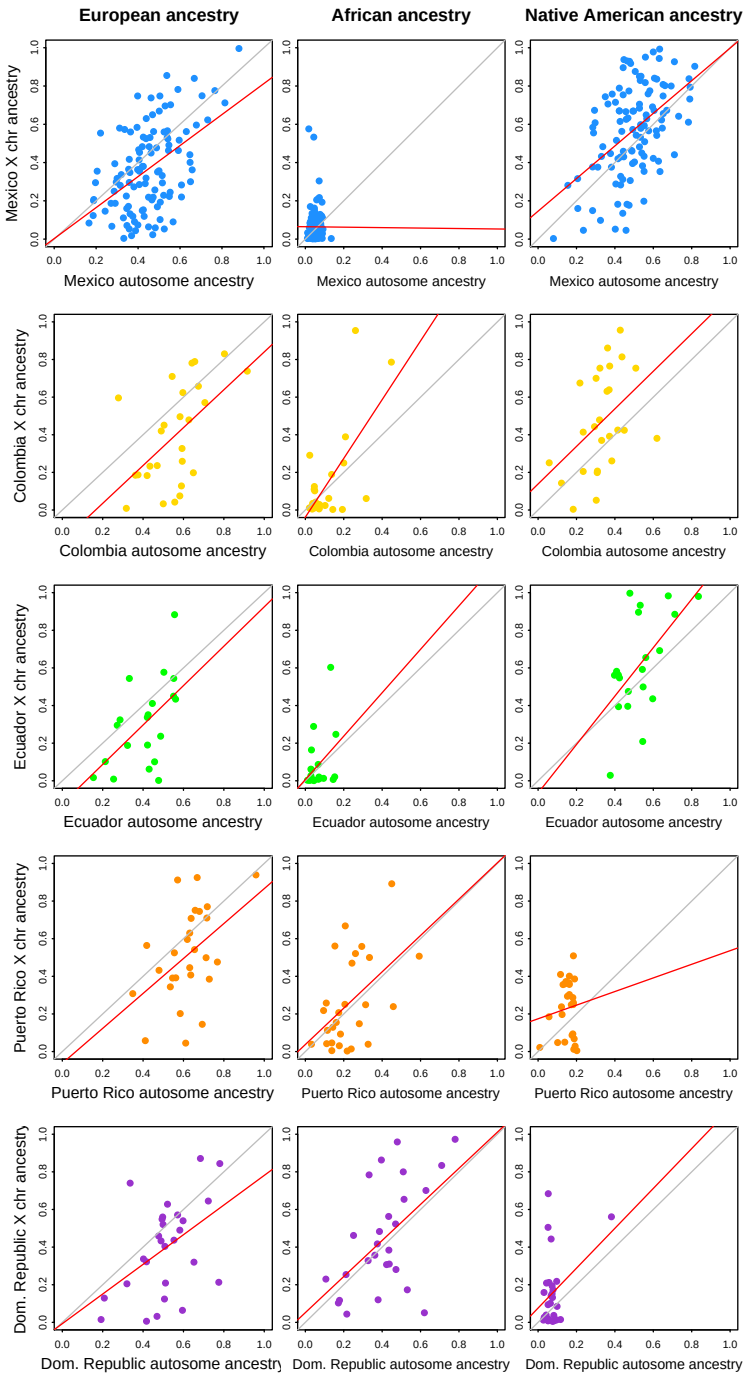

Figure D.4: Individual scatterplots comparing autosomal versus X chromosome ancestry proportions for each population. Expected autosome = X chromosome ancestry lines are show in gray, fitted linear regression lines in red.

BIBLIOGRAPHY

[Adeyemo et al., 2005] Adeyemo, A. A., Chen, G., Chen, Y., and Rotimi, C. (2005). Genetic structure in four west african population groups. *BMC Genet*, 6:38.

[Altshuler et al., 2005] Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., and Donnelly, P. a. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

[Auton et al., 2009] Auton, A., Bryc, K., Boyko, A. R., Lohmueller, K. E., Novembre, J., Reynolds, A., Indap, A., Wright, M. H., Degenhardt, J. D., Gutenkunst, R. N., King, K. S., Nelson, M. R., and Bustamante, C. D. (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res*, 19(5):795–803.

[Bansal et al., 2007] Bansal, V., Bashir, A., and Bafna, V. (2007). Evidence for large inversion polymorphisms in the human genome from hapmap data. *Genome Res.*, 17(2):219–230.

[Barbujani and Goldstein, 2004] Barbujani, G. and Goldstein, D. (2004). Africans and Asians abroad: genetic diversity in Europe. *Annu Rev Genomics Hum Genet*, 5:119–50.

[Basu et al., 2003] Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., Dey, B., Roy, M., Roy, B., Bhattacharyya, N., et al. (2003). Ethnic India: A Genomic View, With Special Reference to Peopling and Structure. *Genome Research*, 13(10):2277–2290.

[Bauchet et al., 2007] Bauchet, M., McEvoy, B., Pearson, L. N., Quillen, E. E., Sarkisian, T., Hovhannesyan, K., Deka, R., Bradley, D. G., and Shriver, M. D. (2007). Measuring european population stratification with microarray genotype data. *Am J Hum Genet*, 80(5):948–956.

[Bosch et al., 2000] Bosch, E., Calafell, F., Perez-Lezaun, A., Clarimon, J., Comas, D., Mateu, E., Martinez-Arias, R., Morera, B., Brakez, Z., Akhayat, O., Sefiani, A., Hariti, G., Cambon-Thomsen, A., and Bertranpetit, J. (2000). Genetic structure of north-west africa revealed by str analysis. *Eur J Hum Genet*, 8(5):360–366.

[Bosch et al., 2002] Bosch, E., Lee, A. C., Calafell, F., Arroyo, E., Henneman, P., de Knijff, P., and Jobling, M. A. (2002). High resolution y chromosome typing: 19 strs amplified in three multiplex reactions. *Forensic Sci Int*, 125(1):42–51.

[Browning and Browning, 2007] Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81(5):1084–1097.

[Bryc et al., 2010] Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S. A., and Bustamante, C. D. (2010). Genome-wide patterns of population structure and admixture in west africans and african americans. *Proc Natl Acad Sci U S A*, 107(2):786–791.

[Campbell and Tishkoff, 2008] Campbell, M. C. and Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet*, 9:403–33.

[Carvajal-Carmona et al., 2003] Carvajal-Carmona, L. G., Ophoff, R., Service, S., Hartiala, J., Molina, J., Leon, P., Ospina, J., Bedoya, G., Freimer, N., and Ruiz-Linares, A. (2003). Genetic demography of antioquia (colombia) and the central valley of costa rica. *Hum Genet*, 112(5-6):534–41.

[Cavalli-Sforza and Bodmer, 1971] Cavalli-Sforza, L. L. and Bodmer, W. F. (1971). *The genetics of human populations*. A Series of books in biology. W. H. Freeman, San Francisco.

[Chen et al., 2000] Chen, J., Iannone, M. A., Li, M. S., Taylor, J. D., Rivers, P., Nelsen, A. J., Slentz-Kesler, K. A., Roses, A., and Weiner, M. P. (2000). A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. *Genome Res*, 10(4):549–557.

[Chikhi et al., 2002] Chikhi, L., Nichols, R. A., Barbujani, G., and Beaumont, M. A. (2002). Y genetic data support the neolithic demic diffusion model. *Proc Natl Acad Sci U S A*, 99(17):11008–11013.

[Choudhry et al., 2006] Choudhry, S., Burchard, E. G., Borrell, L. N., Tang, H., Gomez, I., Naqvi, M., Nazario, S., Torres, A., Casal, J., Martinez-Cruzado, J. C., Ziv, E., Avila, P. C., Rodriguez-Cintron, W., and Risch, N. J. (2006). Ancestry-environment interactions and asthma risk among puerto ricans. *Am J Respir Crit Care Med*, 174(10):1088–93.

[Choudhry et al., 2008] Choudhry, S., Taub, M., Mei, R., Rodriguez-Santana, J., Rodriguez-Cintron, W., Shriver, M. D., Ziv, E., Risch, N. J., and Burchard,

E. G. (2008). Genome-wide screen for asthma in puerto ricans: evidence for association with 5q23 region. *Hum Genet*, 123(5):455–68.

[Clark et al., 2005] Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res*, 15(11):1496–1502.

[Conrad et al., 2006] Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., and Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet*, 38(11):1251–60.

[Coop et al., 2008] Coop, G., Wen, X., Ober, C., Pritchard, J. K., and Przeworski, M. (2008). High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 319(5868):1395–8.

[Dipierri et al., 1998] Dipierri, J. E., Alfaro, E., Martínez-Marignac, V. L., Bailliet, G., Bravi, C. M., Cejas, S., and Bianchi, N. O. (1998). Paternal directional mating in two amerindian subpopulations located at different altitudes in northwestern argentina. *Hum Biol*, 70(6):1001–10.

[Durbin et al., 1999] Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1999). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

[Ehret, 2001] Ehret, C. (2001). Bantu expansions: re-envisioning a central problem of early african history. *International journal of African historical studies*, pages 5–41.

[Emeneau and Burrow, 1962] Emeneau, M. and Burrow, T. (1962). *Dravidian borrowings from Indo-Aryan*. University of California Press, Berkeley; Los Angeles.

[Falush et al., 2003] Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587. Comparative Study.

[Fejerman et al., 2008] Fejerman, L., John, E. M., Huntsman, S., Beckman, K., Choudhry, S., Perez-Stable, E., Burchard, E. G., and Ziv, E. (2008). Genetic ancestry and risk of breast cancer among u.s. latinas. *Cancer Res*, 68(23):9723–8.

[Firmann et al., 2008] Firmann, M., Mayor, V., Vidal, P. M., Bochud, M., Pecoud, A., Hayoz, D., Paccaud, F., Preisig, M., Song, K. S., Yuan, X., Danoff, T. M., Stirnadel, H. A., Waterworth, D., Mooser, V., Waeber, G., and Vollenweider, P. (2008). The colaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovasc Disord*, 8:6.

[Frazer et al., 2007] Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C.,

McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–61.

[Garrigan et al., 2007] Garrigan, D., Kingan, S. B., Pilkington, M. M., Wilder, J. A., Cox, M. P., Soodyall, H., Strassmann, B., Destro-Bisol, G., de Knijff, P., Novelletto, A., Friedlaender, J., and Hammer, M. F. (2007). Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, x and y chromosome resequencing data. *Genetics*, 177(4):2195–207.

[González-Andrade et al., 2007] González-Andrade, F., Sánchez, D., González-Solórzano, J., Gascón, S., and Martínez-Jarreta, B. (2007). Sex-specific genetic admixture of mestizos, amerindian kichwas, and afro-ecuadorans from ecuador. *Hum Biol*, 79(1):51–77.

[González Burchard et al., 2005] González Burchard, E., Borrell, L. N., Choudhry, S., Naqvi, M., Tsai, H.-J., Rodriguez-Santana, J. R., Chapela, R., Rogers, S. D., Mei, R., Rodriguez-Cintron, W., Arena, J. F., Kittles, R., Perez-Stable, E. J., Ziv, E., and Risch, N. (2005). Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. *Am J Public Health*, 95(12):2161–8.

[Green et al., 2000] Green, L. D., Derr, J. N., and Knight, A. (2000). mtdna affinities of the peoples of north-central mexico. *Am J Hum Genet*, 66(3):989–98.

[Hall, 2005] Hall, G. M. (2005). *Slavery and African ethnicities in the Americas: restoring the links*. University of North Carolina Press, Chapel Hill.

[Hayes et al., 2007] Hayes, M. G., Pluzhnikov, A., Miyake, K., Sun, Y., Ng, M. C. Y., Roe, C. A., Below, J. E., Nicolae, R. I., Konkashbaev, A., Bell, G. I., Cox, N. J., and Hanis, C. L. (2007). Identification of type 2 diabetes genes in mexican americans through genome-wide association studies. *Diabetes*, 56(12):3033–44.

[Hellenthal et al., 2008] Hellenthal, G., Auton, A., and Falush, D. (2008). Inferring human colonization history using a copying model. *PLoS Genet*, 4(5):e1000078.

[Hinds et al., 2004] Hinds, D. A., Stokowski, R. P., Patil, N., Konvicka, K., Kershenobich, D., Cox, D. R., and Ballinger, D. G. (2004). Matching strategies for genetic association studies in structured populations. *Am J Hum Genet*, 74(2):317–325.

[Hinds et al., 2005] Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. (2005). Whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–9.

[Hudson, 2002] Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–8.

[Hughes et al., 2008] Hughes, A. L., Welch, R., Puri, V., Matthews, C., Haque, K., Chanock, S. J., and Yeager, M. (2008). Genome-wide snp typing reveals signatures of population history. *Genomics*, 92(1):1–8.

[Jakobsson et al., 2008] Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., and Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003.

[Johnson, 2008] Johnson, J. A. (2008). Ethnic differences in cardiovascular drug response: potential contribution of pharmacogenetics. *Circulation*, 118(13):1383–93.

[Kashyap et al., 2006] Kashyap, V., Guha, S., Sitalaximi, T., Bindu, G., Hasnain, S., and Trivedi, R. (2006). Genetic structure of Indian populations based on fifteen autosomal microsatellite loci. *BMC Genetics*, 7:28.

[Keinan et al., 2007] Keinan, A., Mullikin, J. C., Patterson, N., and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in east asians than in europeans. *Nat Genet*, 39(10):1251–5.

[Kidd et al., 2008] Kidd, J., Cooper, G., Donahue, W., Hayden, H., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453:56–64.

[Klieman, 2003] Klieman, K. A. (2003). *"The Pygmies were our compass": Bantu and Batwa in the history of west central Africa, early times to c. 1900 C.E.* Heinemann, Portsmouth, NH.

[Kong et al., 2002] Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nat Genet*, 31(3):241–7.

[Kooner et al., 2008] Kooner, J. S., Chambers, J. C., Aguilar-Salinas, C. A., Hinds, D. A., Hyde, C. L., Warnes, G. R., Gomez Perez, F. J., Frazer, K. A., Elliott, P., Scott, J., Milos, P. M., Cox, D. R., and Thompson, J. F. (2008). Genome-wide scan identifies variation in mlxipl associated with plasma triglycerides. *Nat Genet*, 40(2):149–151.

[Lai et al., 2009] Lai, C.-Q., Tucker, K. L., Choudhry, S., Parnell, L. D., Mattei, J., García-Bailo, B., Beckman, K., Burchard, E. G., and Ordovás, J. M. (2009). Population admixture associated with disease prevalence in the boston puerto rican health study. *Hum Genet*, 125(2):199–209.

[Lander and Schork, 1994] Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, 265(5181):2037–2048.

[Lao et al., 2008] Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L. A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A. G., Gieger, C., Wichmann, H.-E., Rüther, A., Schreiber, S., Becker, C., Nürnberg, P., Nelson, M. R., Krawczak, M., and Kayser, M. (2008). Correlation between genetic and geographic structure in europe. *Curr Biol*, 18(16):1241–8.

[Li et al., 2008] Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., and Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–4.

[Li et al., 2006] Li, L., Ho, S., Chen, C., Wei, C., Wong, W., Li, L., Hung, S., Chung, W., Pan, W., Lee, M., et al. (2006). Long contiguous stretches of homozygosity in the human genome. *Hum Mutat*, 27:1115–1121.

[Liang et al., 2007] Liang, L., Zöllner, S., and Abecasis, G. R. R. (2007). GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23(12):1565–1567.

[Lind et al., 2007] Lind, J. M., Hutcheson-Dilks, H. B., Williams, S. M., Moore, J. H., Essex, M., Ruiz-Pesini, E., Wallace, D. C., Tishkoff, S. A., O'Brien, S. J., and Smith, M. W. (2007). Elevated male european and female african contributions to the genomes of african american individuals. *Hum Genet*, 120(5):713–22.

[Lohmueller et al., 2008] Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., Clark, A. G., and Bustamante, C. D. (2008). Proportionally more deleterious genetic variation in european than in african populations. *Nature*, 451(7181):994–997.

[Luca et al., 2008] Luca, D., Ringquist, S., Klei, L., Lee, A. B., Gieger, C., Wichmann, H.-E., Schreiber, S., Krawczak, M., Lu, Y., Styche, A., Devlin, B., Roeder, K., and Trucco, M. (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am J Hum Genet*, 82(2):453–463.

[Ma et al., 2005] Ma, L., Marmor, M., Zhong, P., Ewane, L., Su, B., and Nyambi, P. (2005). Distribution of ccr2-64i and sdf1-3'a alleles and hiv status in 7 ethnic populations of cameroon. *J Acquir Immune Defic Syndr*, 40(1):89–95.

[Mailman et al., 2007] Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z. Y., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., Beck, J., Kimelman, M., Shevelev, S., Preuss, D., Yaschenko, E., Graeff, A., Ostell, J., and Sherry, S. T. (2007). The ncbi dbgap database of genotypes and phenotypes. *Nat Genet*, 39(10):1181–1186.

[Manolio et al., 2007] Manolio, T. A., Rodriguez, L. L., Brooks, L., Abecasis, G., Ballinger, D., Daly, M., Donnelly, P., Faraone, S. V., Frazer, K., Gabriel, S., Gejman, P., Guttmacher, A., Harris, E. L., Insel, T., Kelsoe, J. R., Lander, E., McCowin, N., Mailman, M. D., Nabel, E., Ostell, J., Pugh, E., Sherry, S., Sullivan, P. F., Thompson, J. F., Warram, J., Wholley, D., Milos, P. M., and Collins, F. S. (2007). New models of collaboration in genome-wide association studies: the genetic association information network. *Nat Genet*, 39(9):1045–1051.

[Mao et al., 2007] Mao, X., Bigham, A. W., Mei, R., Gutierrez, G., Weiss, K. M., Brutsaert, T. D., Leon-Velarde, F., Moore, L. G., Vargas, E., McKeigue, P. M.,

Shriver, M. D., and Parra, E. J. (2007). A genomewide admixture mapping panel for hispanic/latino populations. *Am J Hum Genet*, 80(6):1171–8.

[Marchini et al., 2007] Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913.

[Marrero et al., 2007] Marrero, A. R., Bravi, C., Stuart, S., Long, J. C., Pereira das Neves Leite, F., Kommers, T., Carvalho, C. M. B., Pena, S. D. J., Ruiz-Linares, A., Salzano, F. M., and Cátira Bortolini, M. (2007). Pre- and post-columbian gene and cultural continuity: the case of the gaucho from southern brazil. *Hum Hered*, 64(3):160–71.

[Martinez-Marignac et al., 2007] Martinez-Marignac, V. L., Valladares, A., Cameron, E., Chan, A., Perera, A., Globus-Goldberg, R., Wacher, N., Kumate, J., McKeigue, P., O'Donnell, D., Shriver, M. D., Cruz, M., and Parra, E. J. (2007). Admixture in mexico city: implications for admixture mapping of type 2 diabetes genetic risk factors. *Hum Genet*, 120(6):807–19.

[McCarthy et al., 2008] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369.

[McEvoy et al., 2009] McEvoy, B. P., Montgomery, G. W., McRae, A. F., Ripatti, S., Perola, M., Spector, T. D., Cherkas, L., Ahmadi, K. R., Boomsma, D., Willemsen, G., Hottenga, J. J., Pedersen, N. L., Magnusson, P. K. E., Kyvik, K. O., Christensen, K., Kaprio, J., Heikkilä, K., Palotie, A., Widen, E., Muilu, J., Syvänen, A.-C., Liljedahl, U., Hardiman, O., Cronin, S., Peltonen, L., Martin, N. G., and Visscher, P. M. (2009). Geographical structure and differential natural selection among north european populations. *Genome Res*, 19(5):804–14.

[Mendizabal et al., 2008] Mendizabal, I., Sandoval, K., Berniell-Lee, G., Calafell, F., Salas, A., Martínez-Fuentes, A., and Comas, D. (2008). Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in cuba. *BMC Evol Biol*, 8:213.

[Myers et al., 2006] Myers, S., Spencer, C. C., Auton, A., Bottolo, L., Freeman, C., Donnelly, P., and McVean, G. (2006). The distribution and causes of meiotic recombination in the human genome. *Biochem Soc Trans*, 34(Pt 4):526–30.

[Nelis et al., 2009] Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S., Piskácková, T., Balascák, I., Peltonen, L., Jakkula, E., Rehnström, K., Lathrop, M., Heath, S., Galan, P., Schreiber, S., Meitinger, T., Pfeufer, A., Wichmann, H.-E., Melegh, B., Polgár, N., Toniolo, D., Gasparini, P., D'Adamo, P., Klovins, J., Nikitina-Zake, L., Kucinskas, V., Kasnauskiene, J., Lubinski, J., Debniak, T., Limborska, S., Khrunin, A., Estivill, X., Rabionet, R., Marsal, S., Julià, A., Antonarakis, S. E., Deutsch, S., Borel, C., Attar, H., Gagnebin, M., Macek, M., Krawczak, M., Remm, M., and Metspalu, A. (2009). Genetic structure of europeans: a view from the north-east. *PLoS One*, 4(5):e5472.

[Nelson et al., 2008] Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., Briley, L. P., Maruyama, Y., Waterworth, D. M., Waeber, G., Vollenweider, P., Oksenberg, J. R., Hauser, S. L., Stirnadel, H. A., Kooner, J. S., Chambers, J. C., Jones, B., Mooser, V., Bustamante, C. D., Roses, A. D., Burns, D. K., Ehm, M. G., and Lai, E. H. (2008). The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet*, 83(3):347–358.

[Nielsen et al., 2005] Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res*, 15(11):1566–75.

[Novembre et al., 2008] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. (2008). Genes mirror geography within europe. *Nature*, 456(7218):98–101.

[Novembre and Stephens, 2008] Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*, 40(5):646–649.

[Oksenberg et al., 2004] Oksenberg, J. R., Barcellos, L. F., Cree, B. A. C., Baranzini, S. E., Bugawan, T. L., Khan, O., Lincoln, R. R., Swerdlin, A., Mignot, E., Lin, L., Goodin, D., Erlich, H. A., Schmidt, S., Thomson, G., Reich, D. E., Pericak-Vance, M. A., Haines, J. L., and Hauser, S. L. (2004). Mapping multiple sclerosis susceptibility to the hla-dr locus in african americans. *Am J Hum Genet*, 74(1):160–167.

[on Harmonization, 2008] on Harmonization, I. C. (2008). Guidance on e15 pharmacogenomics definitions and sample coding; availability. notice. *Fed Regist*, 73(68):19074–19076.

[Parra et al., 2001] Parra, E. J., Kittles, R. A., Argyropoulos, G., Pfaff, C. L., Hiester, K., Bonilla, C., Sylvester, N., Parrish-Gause, D., Garvey, W. T., Jin, L., McKeigue, P. M., Kamboh, M. I., Ferrell, R. E., Pollitzer, W. S., and Shriver, M. D. (2001). Ancestral proportions and admixture dynamics in geographically defined african americans living in south carolina. *Am J Phys Anthropol*, 114(1):18–29.

[Parra et al., 1998] Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., Forrester, T., Allison, D. B., Deka, R., Ferrell, R. E., and Shriver, M. D. (1998). Estimating african american admixture proportions by use of population-specific alleles. *Am J Hum Genet*, 63(6):1839–51.

[Paschou et al., 2007] Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W., and Drineas, P. (2007). Pca-correlated snps for structure identification in worldwide human populations. *PLoS Genet*, 3(9):1672–86.

[Patin et al., 2009] Patin, E., Laval, G., Barreiro, L. B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K. K., Kidd, J. R., Van der Veen, L., Hombert, J.-M., Gessain, A., Froment, A., Bahuchet, S., Heyer, E., and Quintana-Murci, L. (2009). Inferring the demographic history of african farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet*, 5(4):e1000448.

[Patterson et al., 2004] Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., Daly, M. J., and Reich, D. (2004). Methods for high-density admixture mapping of disease genes. *Am J Hum Genet*, 74(5):979–1000.

[Patterson et al., 2006] Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190.

[Price et al., 2007] Price, A., Patterson, N., Yu, F., Cox, D., Waliszewska, A., McDonald, G., Tandon, A., Schirmer, C., Neubauer, J., Bedoya, G., et al. (2007). A Genomewide Admixture Map for Latino Populations. *The American Journal of Human Genetics*, 80(6):1024–1036.

[Price et al., 2006] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–909.

[Pritchard et al., 2000] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59.

[Purcell et al., 2007] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75.

[Rando et al., 1998] Rando, J. C., Pinto, F., Gonzalez, A. M., Hernandez, M., Larruga, J. M., Cabrera, V. M., and Bandelt, H. J. (1998). Mitochondrial dna analysis of northwest african populations reveals genetic exchanges with european, near-eastern, and sub-saharan populations. *Ann Hum Genet*, 62(Pt 6):531–550.

[Reddy, 2007] Reddy, G. (2007). `http://commons.wikimedia.org/wiki/Image:India-states-numbered.svg`.

[Redon et al., 2006] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–54.

[Reed and Tishkoff, 2006] Reed, F. A. and Tishkoff, S. A. (2006). African human diversity, origins and migrations. *Curr Opin Genet Dev*, 16(6):597–605.

[Reed, 1969] Reed, T. E. (1969). Caucasian genes in american negroes. *Science*, 165(895):762–8.

[Reich et al., 2005] Reich, D., Patterson, N., De Jager, P. L., McDonald, G. J., Waliszewska, A., Tandon, A., Lincoln, R. R., DeLoa, C., Fruhan, S. A., Cabre, P., Bera, O., Semana, G., Kelly, M. A., Francis, D. A., Ardlie, K., Khan, O., Cree, B. A. C., Hauser, S. L., Oksenberg, J. R., and Hafler, D. A. (2005). A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat Genet*, 37(10):1113–8.

[Reiner et al., 2005] Reiner, A. P., Ziv, E., Lind, D. L., Nievergelt, C. M., Schork, N. J., Cummings, S. R., Phong, A., Burchard, E. G., Harris, T. B., Psaty, B. M.,

and Kwok, P.-Y. (2005). Population structure, admixture, and aging-related phenotypes in african american adults: the cardiovascular health study. *Am J Hum Genet*, 76(3):463–77.

[Rosenberg, 2004] Rosenberg, N. (2004). DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes*, 4(1):137–138.

[Rosenberg et al., 2006] Rosenberg, N., Mahajan, S., Gonzalez-Quevedo, C., Blum, M., Nino-Rosales, L., Ninis, V., Das, P., Hegde, M., Molinari, L., Zapata, G., et al. (2006). Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet*, 2(12):e215.

[Rosenberg et al., 2002] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602):2381–2385.

[Sabeti et al., 2007] Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., and Consortium, T. I. H. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–8.

[Salari et al., 2005] Salari, K., Choudhry, S., Tang, H., Naqvi, M., Lind, D., Avila, P., Coyle, N., Ung, N., Nazario, S., Casal, J., et al. (2005). Genetic Admixture and Asthma-Related Phenotypes in Mexican American and Puerto Rican Asthmatics. *Genetic Epidemiology*, 29(1):76.

[Salas et al., 2005] Salas, A., Richards, M., Lareu, M.-V., Sobrino, B., Silva, S., Matamoros, M., Macaulay, V., and Carracedo, A. (2005). Shipwrecks and founder effects: divergent demographic histories reflected in caribbean mtdna. *Am J Phys Anthropol*, 128(4):855–60.

[Sankararaman et al., 2008] Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. *Am J Hum Genet*, 82(2):290–303.

[Sans, 2000] Sans, M. (2000). Admixture studies in latin america: from the 20th to the 21st century. *Hum Biol*, 72(1):155–77.

[Sans et al., 2002] Sans, M., Weimer, T. A., Franco, M. H. L. P., Salzano, F. M., Bentancor, N., Alvarez, I., Bianchi, N. O., and Chakraborty, R. (2002). Unequal contributions of male and female gene pools from parental populations in

the african descendants of the city of melo, uruguay. *Am J Phys Anthropol*, 118(1):33–44.

[Schaffner et al., 2005] Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*, 15(11):1576–1583.

[Scott et al., 2007] Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., Prokunina-Olsson, L., Ding, C.-J., Swift, A. J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.-Y., Conneely, K. N., Riebow, N. L., Sprau, A. G., Tong, M., White, P. P., Hetrick, K. N., Barnhart, M. W., Bark, C. W., Goldstein, J. L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T. A., Watanabe, R. M., Valle, T. T., Kinnunen, L., Abecasis, G. R., Pugh, E. W., Doheny, K. F., Bergman, R. N., Tuomilehto, J., Collins, F. S., and Boehnke, M. (2007). A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345.

[Seldin et al., 2007] Seldin, M. F., Tian, C., Shigeta, R., Scherbarth, H. R., Silva, G., Belmont, J. W., Kittles, R., Gamron, S., Allevi, A., Palatnik, S. A., Alvarellos, A., Paira, S., Caprarulo, C., Guillerón, C., Catoggio, L. J., Prigione, C., Berbotto, G. A., García, M. A., Perandones, C. E., Pons-Estel, B. A., and Alarcon-Riquelme, M. E. (2007). Argentine population genetic structure: large variance in amerindian contribution. *Am J Phys Anthropol*, 132(3):455–62.

[Silva-Zolezzi et al., 2009] Silva-Zolezzi, I., Hidalgo-Miranda, A., Estrada-Gil, J., Fernandez-Lopez, J. C., Uribe-Figueroa, L., Contreras, A., Balam-Ortiz, E., del Bosque-Plata, L., Velazquez-Fernandez, D., Lara, C., Goya, R., Hernandez-Lemus, E., Davila, C., Barrientos, E., March, S., and Jimenez-Sanchez, G. (2009). Analysis of genomic diversity in mexican mestizo populations to develop genomic medicine in mexico. *Proc Natl Acad Sci U S A*, 106(21):8611–6.

[Simoni et al., 2000] Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J., and Barbujani, G. (2000). Geographic Patterns of mtDNA Diversity in Europe. *The American Journal of Human Genetics*, 66(1):262–278.

[Sirugo et al., 2008] Sirugo, G., Hennig, B. J., Adeyemo, A. A., Matimba, A., Newport, M. J., Ibrahim, M. E., Ryckman, K. K., Tacconelli, A., Mariani-Costantini, R., Novelli, G., Soodyall, H., Rotimi, C. N., Ramesar, R. S., Tishkoff, S. A., and Williams, S. M. (2008). Genetic studies of african pop-

ulations: an overview on disease susceptibility and response to vaccines and therapeutics. *Hum Genet*, 123(6):557–98.

[Smith et al., 2001] Smith, M. W., Lautenberger, J. A., Shin, H. D., Chretien, J. P., Shrestha, S., Gilbert, D. A., and O'Brien, S. J. (2001). Markers for mapping by admixture linkage disequilibrium in african american and hispanic populations. *Am J Hum Genet*, 69(5):1080–94.

[Smith et al., 2004] Smith, M. W., Patterson, N., Lautenberger, J. A., Truelove, A. L., McDonald, G. J., Waliszewska, A., Kessing, B. D., Malasky, M. J., Scafe, C., Le, E., De Jager, P. L., Mignault, A. A., Yi, Z., De The, G., Essex, M., Sankale, J.-L., Moore, J. H., Poku, K., Phair, J. P., Goedert, J. J., Vlahov, D., Williams, S. M., Tishkoff, S. A., Winkler, C. A., De La Vega, F. M., Woodage, T., Sninsky, J. J., Hafler, D. A., Altshuler, D., Gilbert, D. A., O'Brien, S. J., and Reich, D. (2004). A high-density admixture map for disease gene discovery in african americans. *Am J Hum Genet*, 74(5):1001–13.

[Stefansson et al., 2005] Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., Desnica, N., Hicks, A., Gylfason, A., Gudbjartsson, D. F., Jonsdottir, G. M., Sainz, J., Agnarsson, K., Birgisdottir, B., Ghosh, S., Olafsdottir, A., Cazier, J. B., Kristjansson, K., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., Kong, A., and Stefansson, K. (2005). A common inversion under selection in europeans. *Nat Genet*, 37(2):129–137.

[Sulem et al., 2008] Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S. A., Palsson, A., Thorleifsson, G., Pálsson, S., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., Vermeulen, S. H., Goldstein, A. M., Tucker, M. A., Kiemeney, L. A., Olafsson, J. H., Gulcher, J., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2008). Two newly identified genetic determinants of pigmentation in europeans. *Nat Genet*, 40(7):835–7.

[Tang et al., 2007] Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E., and Risch, N. (2007). Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans. *The American Journal of Human Genetics*, 81(3):626–633.

[Tang et al., 2005] Tang, H., Peng, J., Wang, P., and Risch, N. J. (2005). Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol*, 28(4):289–301.

[The Indian Genome Variation Consortium, 2008] The Indian Genome Variation Consortium (2008). Genetic landscape of the people of India: a canvas for disease gene exploration. *J Genet*, 87:3–20.

[Thiers et al., 2008] Thiers, F., Sinskey, A., and Berndt, E. (2008). Trends in the globalization of clinical trials. *Nature Reviews Drug Discovery*, 7(1):13–14.

[Tian et al., 2007] Tian, C., Hinds, D., Shigeta, R., Adler, S., Lee, A., Pahl, M., Silva, G., Belmont, J., Hanson, R., Knowler, W., et al. (2007). A Genomewide Single-Nucleotide–Polymorphism Panel for Mexican American Admixture Mapping. *The American Journal of Human Genetics*, 80(6):1014–1023.

[Tian et al., 2008] Tian, C., Plenge, R. M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., Klareskog, L., Pulver, A. E., Qi, L., Gregersen, P. K., and Seldin, M. F. (2008). Analysis and application of european genetic substructure using 300 k snp information. *PLoS Genet*, 4(1):e4.

[Tishkoff et al., 1996] Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonné-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., and Krings, M. (1996). Global patterns of linkage disequilibrium at the cd4 locus and modern human origins. *Science*, 271(5254):1380–7.

[Tishkoff and Kidd, 2004] Tishkoff, S. A. and Kidd, K. K. (2004). Implications of biogeography of human populations for 'race' and medicine. *Nat Genet*, 36(11 Suppl):S21–7.

[Tishkoff et al., 2009] Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., and Williams, S. M. (2009). The genetic structure and history of africans and african americans. *Science*, 324(5930):1035–44.

[Tishkoff et al., 2007] Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., and Deloukas, P. (2007). Convergent adaptation of human lactase persistence in africa and europe. *Nat Genet*, 39(1):31–40.

[Tishkoff and Verrelli, 2003] Tishkoff, S. A. and Verrelli, B. C. (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet*, 4:293–340.

[Tishkoff and Williams, 2002] Tishkoff, S. A. and Williams, S. M. (2002). Genetic analysis of african populations: human evolution and complex disease. *Nat Rev Genet*, 3(8):611–21.

[Tuzun et al., 2005] Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, A. V., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M. V., and Eichler, E. E. (2005). Fine-scale structural variation of the human genome. *Nat Genet*, 37(7):727–732.

[Voight et al., 2006] Voight, B., Kudaravalli, S., Wen, X., and Pritchard, J. (2006). A map of recent positive selection in the human genome. *PLoS Biol*, 4(3):e72.

[Wang et al., 2008] Wang, S., Ray, N., Rojas, W., Parra, M. V., Bedoya, G., Gallo, C., Poletti, G., Mazzotti, G., Hill, K., Hurtado, A. M., Camrena, B., Nicolini, H., Klitz, W., Barrantes, R., Molina, J. A., Freimer, N. B., Bortolini, M. C., Salzano, F. M., Petzl-Erler, M. L., Tsuneto, L. T., Dipierri, J. E., Alfaro, E. L., Bailliet, G., Bianchi, N. O., Llop, E., Rothhammer, F., Excoffier, L., and Ruiz-Linares, A. (2008). Geographic patterns of genome admixture in latin american mestizos. *PLoS Genet*, 4(3):e1000037.

[Weir and Cockerham, 1984] Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population-structure. *Evolution*, 38(6):1358–1370.

[Williamson et al., 2000] Williamson, C., Loubser, S. A., Brice, B., Joubert, G., Smit, T., Thomas, R., Visagie, M., Cooper, M., and van der Ryst, E. (2000). Allelic frequencies of host genetic variants influencing susceptibility to hiv-1 infection and disease in south african populations. *AIDS*, 14(4):449–51.

[Williamson et al., 2007] Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., and Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. *PLoS Genet*, 3(6):e90.

[Willis and Whittaker, 2000] Willis, K. and Whittaker, R. (2000). The refugial debate. *Science*, 287(5457):1406–1407.

[Workman et al., 1963] Workman, P. L., Blumberg, B. S., and Cooper, A. J. (1963). Selection, gene migration and polymorphic stability in a u. s. white and negro population. *Am J Hum Genet*, 15:429–37.

[WTCCC, 2007] WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78.

[Xu and Jin, 2008] Xu, S. and Jin, L. (2008). A genome-wide analysis of admixture in uyghurs and a high-density admixture map for disease-gene discovery. *Am J Hum Genet*, 83(3):322–36.

[Yu et al., 2006] Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, 38(2):203–208.